Discourse-level features for statistical machine translation

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 6501 (2014)
présentée le 8 décembre 2014
à l'Institut de Recherche Idiap
Faculté Sciences et Techniques pour l'Ingénieur (STI)
Programme Doctoral en Génie Électrique (EDEE)
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Thomas Meyer

soumise le 24 octobre 2014 au jury:

Prof. Jean-Philippe Thiran, président du jury
Prof. Hervé Bourlard, directeur de thèse
Dr. Andrei Popescu-Belis, co-directeur de thèse
Dr. Martin Rajman, rapporteur
Prof. Bonnie Webber, rapporteur
Prof. Martin Volk, rapporteur

Lausanne, EPFL, 2014

Translation is that which transforms everything so that nothing changes.
– Günter Grass

Acknowledgements

Of our elaborate plans, the end
Of everything that stands, the end
No safety or surprise, the end
I'll never look into your eyes...again
Jim Morrison, The End

Writing this thesis has come to an end. I however do not look back to it as consequently as the writer of the above lines does – especially because I was allowed to do research within ideal circumstances.
My due thanks go to my advisors Andrei Popescu-Belis and Hervé Bourlard. Andrei supported my work in every thinkable way and I could not have written this thesis without his constant input and the strive for better results. I am deeply grateful for his help and availability during the past years.
I also would like to acknowledge the Swiss National Science Fundation (SNF) who funded this work through the two Sinergia projects COMTIS and MODERN.
Furthermore, the Idiap Research Institute and all its staff provided an environment of friendliness, flexibility, and enjoyable learning.
I am thankful to my thesis committee – Martin Rajman, Jean-Philippe Thiran, Martin Volk and Bonnie Webber who reviewed the present thesis in great detail and provided feedback that greatly improved the final version.
Special thanks also go to Bonnie Webber, who made it possible that I could visit the University of Edinburgh during an internship for which I will keep all the best memories, both professionally and privately.
Collaborations with the University of Geneva further diversified and enriched the present work. I would like to specially thank Sandrine, Bruno, Andrea, Sharid and Cristina, but also Jacques, Paola, James and Tanja.
At my new workplace, thanks to my colleagues for the warm welcome, and for all collaboration further down the NLP road: Linne, Bruno, Eli, Silvia, Adnan, Daniel and the entire team.
At Idiap, I first of all would like to thank my colleagues in office 207.1: Laurent, André, Manuel, Ivana, Elie, Carl, Roy, Nesli and from other offices: Chris, Rémi, Alexandre, Kenneth, Maryam, David, Flavio, Dimitri. The very many coffee breaks, Friday beers, discussions and group activities would not have been what they were without these very people. Thanks also to König and its high-quality Nespresso machine Capri which spat out about 20 coffees per day during more than four years – it all made us survive.

Acknowledgements

Climber's thanks to Vincent, Carl, Marco, James, Laurent and Petr who helped to keep in shape and who always were a fun party not only on the rocky walls of Martigny. For other socially most enjoyable hours I would like to further thank Phil, Hari, Laurent, Darshan, Leo, Alexandros.
Special thanks go to Majid and Nikos, who provided personal, psychological and scientific support throughout the years and who lived through many a good and difficult hour with me.
Further family and friends deserve my gratitude: Alois, Martin, Peter, Ramon, Sara, Serge, Nathie, Jan, Lukas, Martin, Urs, Lisa for providing every thinkable support: mentally, socially, financially and up to a bed in emergency. Keep the spirit, sisters and brothers.
My deepest, heartmost thanks go to Anja, Braida and Susanna – the three most important women in my life (for different reasons): For their love, encouragement, enlightenment and without whom I would not be who I am today.

Lausanne, January 2015 Th. M.

Abstract

Discourse-level features for statistical machine translation

Machine Translation (MT) has progressed tremendously in the past two decades. The rule-based and interlingua approaches of the 1980s have been superseded by statistical models, which learn the most likely translations from large parallel corpora that became available via the internet. Such resources consist of source language texts that are aligned to human reference translations. System design does not amount anymore to crafting grammatical and syntactical transfer rules, nor does it rely on a semantic representation of the source text's meaning to generate the target text from. Instead, during a training stage, a statistical MT system learns the most likely correspondences and re-ordering of chunks of source words and target words from parallel corpora that have been sentence- and word-aligned. These chunks or 'phrases' are not necessarily linguistically motivated (in terms of sentence constituents, for example). With this procedure and millions of parallel source and target language sentences, systems can generate translations that are intelligible and require minimal post-editing efforts from the human user.

Nevertheless, it has been recognized as early as at the beginning of the 1990s that the statistical MT paradigm may fall short of modeling a number of linguistic phenomena that are established beyond the phrase or sentence level. Research in statistical MT has merely focused on lexical choice and syntactical structures, and has addressed coherence or discourse phenomena explicitly only in the past four years.

When it comes to textual structure or text coherence, the cohesive ties or markers relate sentences and entire paragraphs argumentatively to each other. This text structure has to be rendered appropriately in the target text so that it conveys the same meaning as the source text. The lexical and syntactical means through which these cohesive markers are expressed may diverge considerably between languages. Frequently, these markers include discourse connectives, which are a class of function words such as *although, however, instead, meanwhile*, or *since*, which relate two spans of text to each other, e.g. for temporal ordering, contrast, elaboration or causality. Moreover, to establish the same temporal ordering of the content or events described in a text, the verbal tense, mode and aspect has to be coherently translated so that the reader of the target text can infer the correct sequence and meaning of what is described.

The present thesis proposes methods for integrating textual coherence and discourse features into statistical MT. Rather than trying to store previously translated units in a cache or to model the topic distribution or the lexical consistency of a document, which are other recent

attempts in this direction, we propose to pre-process the source text prior to automatic translation, focusing on two specific discourse phenomena: discourse connectives and verb tenses. Hand-crafted rules are not required in our proposal; instead, machine learning classifiers are implemented that learn to recognize discourse relations or to predict translations of verb tenses. The classifiers are then used to automatically annotate the corresponding word forms in a source text. Similar techniques have been used in recent research work, but most often only for content word disambiguation in MT. To address function words in this manner is a novelty and we have shown that complex features from a long-range context are beneficial for disambiguating connectives and verb tenses.

The contributions of the present thesis are two-fold. Firstly, we have designed new sets of semantically-oriented features and specific classifiers to advance the state of the art in automatic disambiguation or classification of discourse connectives. This remains an open NLP problem in its own right. For this, we profited from our multilingual setting and incorporated features that are based on MT and on the insights we gained from contrastive linguistic analysis of parallel corpora. In their best configurations, our classifiers reach high performances (0.7 to 1.0 F1 score) and can therefore reliably be used to automatically annotate the large corpora needed to train SMT systems. Issues of manual annotation and evaluation of the classifiers are discussed in the thesis, and solutions are provided within new annotation procedures and evaluation metrics. The annotated resources and the disambiguation models have been made available to the community for reproducibility and further research.

As a second contribution, we implemented entire SMT system pipelines that can make use of, and learn from, the (automatically) annotated discourse information to translate these elements more correctly and consequently to generate more coherent target text. A number of methods have been tested for this purpose, from factored translation models used in previous research work, to original methods that make maximum use of the information provided by the classifiers in form of label probability scores that can be used for the translation process as well.

Overall, the thesis confirms that the technique of pairing discourse-level classifiers and statistical MT is a practical and workable solution that leads to global improvements in translation in ranges of 0.2 to 0.5 BLEU score. We additionally performed automatic and manual evaluations of translation quality by comparing translation output from unmodified baseline SMT systems with the output of system variants that were trained on input texts labeled with discourse relations and verb tenses. These evaluations clearly revealed that in terms of connectives and verb tenses, our statistical MT systems improve the translation of these phenomena in ranges of up to 25%, depending on the performance of the automatic classifiers, the data sets and the system configurations.

Keywords: Statistical Machine Translation, Discourse, Discourse Relations, Discourse Connectives, Verb Tenses

Résumé

Utilisation de traits discursifs pour la traduction automatique statistique

La traduction automatique (TA) a progressé énormément pendant les deux dernières décennies. Les approches des années 1980 à base de règles et celles qui utilisent une interlangue ont été remplacées par des modèles statistiques qui apprennent les traductions les plus probables grâce à de grands corpus parallèles disponibles via l'internet. Ces ressources se composent de textes en langue source qui sont alignés avec des traductions de référence produites par des humains. La conception des systèmes de TA, dans cette approche, ne requiert plus l'implémentation de règles de transfert grammatical et syntaxique, ni la représentation sémantique de la signification du texte source à partir duquel le texte cible est généré. Au lieu de cela, un système de TA statistique apprend les correspondances les plus probables et la réorganisation de groupes de mots source et des mots cible à partir des corpus parallèles qui ont été alignés au niveau des mots. Ces groupes de mots ne sont pas nécessairement motivés linguistiquement (en termes de constituants syntaxiques, par exemple). Avec cette procédure, appliquée à des millions de phrases parallèles, les systèmes peuvent générer des traductions qui sont compréhensibles et nécessitent un minimum de post-édition de la part de l'utilisateur humain.

Néanmoins, il a été reconnu dès le début des années 1990 que le paradigme de la TA statistique rencontre des difficultés lorsqu'il s'agit de modéliser un certain nombre de phénomènes linguistiques qui s'établissent en dehors des limites d'une phrase ou d'une clause. La recherche en TA statistique a mis l'accent sur le choix lexical et les structures syntaxiques, et n'a abordé la cohérence ou les phénomènes de discours explicitement que durant les quatre dernières années.

Afin de structurer un texte et d'en assurer la cohérence, les marques de cohésion permettent de connecter les arguments des phrases et des paragraphes. La structure textuelle doit être rendue de manière appropriée dans le texte cible d'une manière, à savoir avec le même sens que dans le texte source. Les moyens lexicaux et syntaxiques par lesquels les marques de cohésion sont exprimées peuvent diverger considérablement entre les langues. Souvent, ces marques comprennent des connecteurs de discours, qui sont une classe de mots de fonction tels que *bien que, cependant, entre-temps, pendant que* ou *depuis*, qui attachent deux clauses ou phrases l'une à l'autre, par exemple pour exprimer la temporalité, le contraste, l'élaboration ou la causalité. En outre, pour établir le même ordre temporel des événements décrits dans un texte, la conjugaison des verbes en termes de temps, mode et aspect doit être traduite de manière cohérente afin que le lecteur du texte cible puisse comprendre correctement leur

séquence.

Cette thèse propose des méthodes pour intégrer des traits qui modélisent la cohérence textuelle et la structure du discours dans la TA statistique. Plutôt que d'essayer de stocker les unités déjà traduites ou de modéliser la distribution des sujets ou la cohérence lexicale d'un document (des questions qui font l'objet d'autres tentatives récentes dans ce sens), nous proposons de prétraiter le texte source avant la traduction automatique, en nous concentrant sur deux phénomènes discursifs spécifiques : les connecteurs de discours et les temps verbaux. Des règles rédigées manuellement ne sont pas nécessaires dans notre proposition ; à leur place, des classifieurs fondés sur l'apprentissage automatique sont mis en œuvre pour apprendre à reconnaître les relations de discours ou à prédire les traductions des temps verbaux. Les classifieurs sont utilisés pour annoter automatiquement les mots correspondants dans le texte source. Des techniques similaires ont été utilisées dans des travaux de recherche récents, mais le plus souvent seulement pour la désambiguïsation des mots de contenu dans la TA. Le traitement des mots de fonction de cette manière représente ainsi une proposition nouvelle, et nous avons pu montrer que des traits complexes dérivés d'un contexte plus étendu étaient bénéfiques pour lever l'ambiguïté des connecteurs et les temps verbaux.

Les contributions de cette thèse s'organisent sur deux axes. D'abord nous utilisons des traits sémantiques et des classifieurs spécifiques pour faire progresser l'état de l'art de la désambiguïsation automatique des connecteurs de discours (qui représente un problème encore ouvert en TAL). Nous avons profité de la problématique multilingue de nos travaux, à travers des traits qui sont basés sur la TA et des analyses linguistiques contrastives de corpus parallèles. Dans leurs meilleures configurations, nos classifieurs atteignent des performances élevées (de 0.7 à 1.0 score F1) et peuvent donc être utilisés de manière fiable pour annoter automatiquement des grandes corpus nécessaires pour entraîner des systèmes TA statistiques. Les problèmes de l'annotation manuelle et de l'évaluation sont discutés également, et des solutions sont proposées avec de nouvelles procédures d'annotation et métriques d'évaluation. Les ressources annotées et les modèles de désambiguïsation ont été mis à la disposition de la communauté pour la reproductibilité de nos recherches.

Comme une seconde contribution, nous avons implémenté des systèmes de TA statistiques complets qui peuvent apprendre et utiliser les informations au niveau du discours (annotées automatiquement ou non) pour traduire ces éléments plus correctement et pour générer des textes cible plus cohérents. Un certain nombre de méthodes ont été testées à cet effet, à partir de modèles de traduction avec facteurs (provenant de travaux de recherche antérieurs), jusqu'à des méthodes originales qui utilisent dans le processus de traduction les étiquettes fournies par les classificateurs ainsi que leurs probabilités.

Dans l'ensemble, la thèse confirme que la classification au niveau du discours, apprise automatiquement, et la TA statistique sont une solution pratique et réalisable qui conduit à des améliorations globales de la traduction de l'ordre de 0.2 à 0.5 sur la métrique BLEU. Nous avons effectué des évaluations automatiques et manuelles de la qualité de la traduction en comparant les résultats des systèmes de TA statistiques non modifiés avec la sortie des systèmes qui ont été entraînés sur des textes d'entrée marqués avec les relations de discours et les temps verbaux. Ces évaluations ont révélé clairement qu'en termes de connecteurs et de

temps verbaux, nos systèmes de TA statistique améliorent la traduction de ces phénomènes jusqu'à 25%, en fonction de la performance des classifieurs automatiques, des données et des configurations du système.

Mots-clés : Traduction Automatique Statistique, Discours, Relations de Discours, Connecteurs de Discours, Temps Verbaux

Zusammenfassung

Diskurs-Features für die statistische maschinelle Übersetzung

Die maschinelle Übersetzung (MÜ) hat in den letzten beiden Dekaden enorme Fortschritte gemacht. Die regel- und Interlingua-basierten Vorgehensweisen der 1980er Jahre wurden durch statistische Modelle abgelöst, welche die wahrscheinlichsten Übersetzungen aus grossen, parallelen Korpora lernen, die übers Internet zugänglich geworden sind. Diese Ressourcen bestehen aus quellsprachlichen Texten, die mit Referenzübersetzungen aligniert sind. Die Implementierung von Übersetzungsystemen muss sich nicht mehr auf syntaktische Transferregeln oder auf die semantische Repräsentation der Quelltextbedeutung verlassen, um den Zieltext generieren zu können. Stattdessen lernt ein statistiches Übersetzungssystem die häufigsten und wahrscheinlichsten Entsprechungen sowie Worstellungen von Quell- und Zielphrasen aus parallelen Korpora, die satz- und wortaligniert sind. Diese Phrasen sind dabei nicht zwingend linguistisch motiviert (im Sinne von Satzkonstituenten). Mit diesem Vorgehen und Millionen von parallelen Quell- und Zielsätzen können die Systeme Übersetzungen generieren, die verständlich sind und nur minimale Nachbearbeitung durch den Benutzer erfordern.

Nichtsdestotrotz wurde bereits zu Beginn der 90er Jahre erkannt, dass das statistische MÜ-Paradigma nicht alle linguistischen Phänomene modellieren kann, insbesondere nicht solche, die über die Phrasen- oder Satzgrenzen hinausgehen. Die MÜ-Forschung hat sich vorerst aber mehr darauf konzentriert, lexikalische Konsistenz oder wohlgeformte Syntaxstrukturen zu erhalten, weshalb textuelle Kohärenz oder Diskursphänomene erst in den letzten vier Jahren in den Fokus gerückt sind. Textuelle Kohärenz etabliert sich über sogenannte Kohäsionspartikel oder -marker, die Sätze und ganze Paragraphen argumentativ verbinden. Diese Textstruktur muss in der Zielsprache entsprechend korrekt wiedergegeben werden, damit der Zieltext die exakte Bedeutung des Quelltexts vermittelt. Die lexikalischen und syntaktischen Mittel, mit denen diese Kohäsionspartikel zum Ausdruck kommen, können sich von Sprache zu Sprache erheblich unterscheiden. Oft beinhalten die Kohäsionspartikel die sogenannten Diskurskonnektoren, welche eine funktionale Wortklasse bilden und zwei Textspannen miteinander verbinden um Temporalität, Kausalität, Weiterführung oder Kontrast zu etablieren (zu ihnen gehören Wörter wie: *obwohl, jedoch, stattdessen, während, in der Zwischenzeit, da, seit*, etc.).

Ferner spielt die korrekte Konjugation von Verben betreffend Tempus, Modus und Aspekt eine grosse Rolle, um dieselbe zeitliche Ordung der im Quelltext beschriebenen Ereignisse bei der Übersetzung im Zieltext wiederzugeben. Die vorliegende Arbeit stellt neue Methoden auf um

Diskurs-Features in die statistische MÜ zu integrieren. Statt wie in anderen Arbeiten zum Thema zu versuchen bereits übersetzte Einheiten zu speichern oder die Themendistribution und die lexikalische Konsistenz eines Dokuments zu modellisieren, wird hier eine Annotation der Diskurskonnektoren und der Verbtempora im Quelltext vorgeschlagen, bevor ein Text zur Übersetzung gelangt. Die manuelle Implementation von Regeln ist hierfür nicht erforderlich; vielmehr kommen Maschinelles Lernen und Klassifikatoren zum Einsatz, die lernen, Diskursrelationen oder Verbtempora im Zieltext vorauszuberechnen.
Ähnliche Methoden wurden bereits für die Wortbedeutungsdesambiguierung in der MÜ angewandt; dies aber meist nur für Inhaltswörter. Es ist ein Novum der vorliegenden Arbeit, dies auf Funktionswörter auszuweiten und es wird gezeigt, dass Features aus einem breiteren Kontext zur Desambiguierung von Konnektoren und Verbtempora hilfreich sein können.
Der Forschungsbeitrag der vorliegenden Arbeit ist zweiteilig: Zum einen werden neue Sets an semantisch orientieren Features vorgeschlagen, um die Performanz der automatischen Desambiguierung von Diskurskonnektoren zu erhöhen. Dies ist ein ungelöster Forschungspunkt der heutigen Computerlinguistik. Beim Finden der Features halfen das mutlinguale Setting, der Einsatz paralleler Korpora und die kontrastive Textanalyse. Mit den besten Konfigurationen erreichen die Klassifikatoren hohe Performanz (F1 scores im Bereich von 0.7 bis 1.0) und können deshalb verlässlich zur automatischen Annotation der Konnektoren in den grossen Textmengen, die für die MÜ nötig sind, eingesetzt werden. Problembereiche der manuellen Annotation werden ebenso diskutiert wie Evaluationsmethoden und Lösungen mit neuen Annotationsmethoden und Evaluationsmetriken. Die annotierten Ressourcen und die Desambiguierungsmodelle sind für die weitere Forschung erhältlich und sorgen für die Nachvollziehbarkeit der Ergebnisse. Der zweite Beitrag besteht aus Implementationen kompletter statistischer MÜ-Systeme, die die (automatisch) annotierten Korpora benützen und entsprechend lernen, die Diskursinformation präziser zu übersetzen. Damit wird der generierte Zieltext kohärenter. Eine ganze Reihe an Methoden zur Integration der Annotationen in die MÜ-Prozesse wurde getestet, von sogenannten Factored Translation Models bis hin zu Modellen, die die von Klassifikatoren gelieferte Information maximal ausnützen und die Distribution der annotierten Relationen in den Daten berücksichtigen.
Die vorliegende Arbeit bestätigt, dass sich Klassifikatoren auf der Diskursebene erfolgreich in MÜ-Systeme einbauen lassen und eine durchführbare Methode darstellen, um globale Verbesserungen in der MÜ-Ausgabe im Bereich von 0.2 bis zu 0.5 BLEU-Punkten zu erhalten. Zusätzlich wurde die Übersetzungsqualität der Systeme mit anderen automatischen und manuellen Metriken evaluiert. Dabei wurden die Ausgaben von unveränderten Basissystemen mit den Ausgaben von modifizierten Systemen verglichen, die darauf trainiert wurden, die Diskursannotation zu berücksichtigen. Diese Evaluationen zeigen klar, dass die Übersetzungsqualtät für Konnektoren und Verbtempora mit Werten von bis zu 25% ansteigt, basierend auf der Performanz der Klassifikatoren, der Qualität der Daten und der Konfiguration der MÜ-Systeme.

Stichwörter: Statistische Maschinelle Übersetzung, Diskurs, Diskursrelationen, Diskurskonnektoren, Verbtempus

Contents

Acknowledgements **v**

Abstract (English / Français / Deutsch) **vii**

List of figures **xvii**

List of tables **xx**

1 Introduction **1**
1.1 Rule-based vs. statistical machine translation 3
1.2 Related research projects 6
1.3 Contributions of the thesis 7

2 Discourse connectives and verb tense in translation **13**
2.1 Discourse connectives in translation 14
2.1.1 Translation problems related to connectives 14
2.1.2 Examples of translation errors 15
2.2 Verb tense in translation 17

3 Related work **21**
3.1 Discourse processing 21
3.1.1 Discourse parsing 21
3.1.2 Disambiguating discourse connectives 23
3.2 Modeling verb tense 24
3.3 Statistical machine translation (SMT) 26
3.3.1 Mathematical definition of phrase-based statistical machine translation 26
3.3.2 SMT models for using linguistic information 27
3.3.3 Verb tense in SMT 31

4 Data, annotation procedures and evaluation metrics **33**
4.1 Data 34
4.1.1 The Penn Discourse Treebank 34
4.1.2 The Europarl corpus 36
4.1.3 Other corpora used for statistical machine translation 38
4.2 Annotation procedures 39

4.2.1 Discourse relations . . . 39
4.2.2 Annotation of verb tense . . . 45
4.3 Evaluation metrics . . . 49

5 Automatically disambiguating discourse connectives 53
5.1 Algorithms . . . 54
5.2 Connective labeling vs. word sense disambiguation . . . 55
5.3 Features for connective labeling . . . 56
5.4 Disambiguation experiments based on the PDTB . . . 60
5.5 Disambiguation experiments based on Europarl . . . 64
5.6 Experiments on large feature and data sets . . . 67
5.6.1 Merging PDTB and Europarl data . . . 67
5.6.2 Feature analysis and selection . . . 69
5.6.3 Significance of connective labeling scores . . . 71
5.6.4 Results on the test sets . . . 72

6 Automatically disambiguating verb tense 75
6.1 Disambiguating narrativity . . . 75
6.2 Automatically predicting French verb tense . . . 77
6.2.1 Features . . . 77
6.2.2 Results . . . 80

7 Statistical machine translation with discourse labels 83
7.1 Oracle experiments . . . 85
7.1.1 SMT with oracle disambiguation of connectives . . . 85
7.1.2 Oracle SMT with verb tense . . . 89
7.2 Phrase table modification . . . 91
7.3 Concatenating labels to word forms . . . 93
7.4 System combination based on labeling confidence . . . 97
7.5 Duplication of training data based on label confidence . . . 99
7.6 Post-editing discourse connectives . . . 101
7.7 Factored Models . . . 102
7.7.1 Factored models with discourse and POS labels . . . 103
7.7.2 Factored models with discourse labels across multiple target languages . . . 105
7.8 SMT with labels for verb tense . . . 108
7.8.1 SMT with narrativity labels . . . 108
7.8.2 SMT with predicted French tense labels . . . 113
7.9 Conclusions on factored translation models . . . 115

8 Statistical machine translation with deletion/insertion of connectives 117
8.1 Semi-automatic corpus analyses for implicitation/explicitation of discourse connectives . . . 117
8.1.1 Implicitation of connectives . . . 118

8.1.2 Explicitation of connectives . . . 125
8.2 Sparse lexical features for SMT . . . 130
8.2.1 SMT tuning with lexical features . . . 130
8.2.2 Data . . . 131
8.2.3 Models . . . 132
8.2.4 Results and discussion . . . 132

9 Conclusions and perspectives **139**
9.1 Conclusions . . . 139
9.2 Perspectives . . . 142

A Appendix **145**

Bibliography **171**

Curriculum Vitae **173**

List of Figures

1.1 Mistranslations at the discourse level 1
1.2 Vauquois pyramid of MT system paradigms 4

2.1 Mistranslation of the English connective *since* in French 15
2.2 Mistranslation of the English connective *while* in German 16
2.3 Equivalent Italian translations of the English connective *even though* 16
2.4 Baseline translation examples of English Simple Past verbs in (non-)narrative context 17
2.5 Examples of translations with English Simple Past verbs by human translators and a baseline SMT system 19
2.6 Example of English/French verb tense divergency with labeled verbs 20

3.1 Example of a Rhetorical Structure Theory discourse tree 22

4.1 Penn Discourse Treebank hierarchy of discourse relations 36
4.2 Translation spotting example with the English connective *since* 41
4.3 English/French word alignment and parsing for inferring French verb tenses . 49

7.1 Examples of English/French training data with discourse connectives with concatenated sense tags 94
7.2 Examples of label probability scores for connectives as output by a MaxEnt disambiguation model 97
7.3 Varying translation performance when combining baseline and discourse-aware English/French SMT systems 98
7.4 Example input for factored translation models with POS tags and labels for connectives 103
7.5 Example input for a factored translation model with narrativity labels for English Simple Past verbs 109
7.6 Example of an improved English/French translation with labeled narrativity . . 112

8.1 Examples of English/French and English/German connective translations with implicitation 118
8.2 English connective dictionary entries with equivalents and paraphrases in French and German 120

8.3 Percentage of implicitation per discourse relation for English/French translation 121
8.4 Percentage of implicitation per discourse relation in English/German translation 122
8.5 Translation examples for the English temporal connectives *while* and *when*, rendered in the French reference as a 'preposition + Verb in Gerund' construction. 123
8.6 Translation example for the English connective *if*, rendered in the German reference as a construction with a sentence-initial verb in conditional mood 125
8.7 Dictionary entry for the FR connective *malgré tout* 127
8.8 Dictionary entry for the DE connective *vielmehr* 127
8.9 Explicitation percentage per discourse relation for English/French translation . 128
8.10 Explicitation percentage per discourse relation for English/German translation 128
8.11 Example of connective explicitation in English/French translation 129
8.12 Example of connective explicitation in English/German translation 129
8.13 Example excerpt of a source word deletion list 133
8.14 Implicitation of discourse connectives in human reference translations, and translations output by a baseline SMT system compared to SMT models with sparse lexcial source word deletion features 134
8.15 Examples of implicitated connective translations for English/German 135

List of Tables

2.1 Distribution of English/French verb tense translations in the Europarl corpus . 20

4.1 Language tags in the Europarl corpus 38
4.2 Improving language tags in the Europarl corpus 38
4.3 Sense clustering for the English connective *while* after translation spotting . . . 42
4.4 List of annotated resources for discourse connectives in English/French 45
4.5 Datasets for English/French verb tense prediction and SMT 50

5.1 Accuracy for the disambiguation of eight English temporal–contrastive connectives 61
5.2 Performance of MaxEnt classifier configurations for English connectives 63
5.3 Disambiguation performance for the connectives *alors que, since, while* 64
5.4 Information gain of features for the French connective *alors que* 65
5.5 Information gain of features for the English connective *since* 66
5.6 Information gain of features for the English connective *while* 66
5.7 Disambiguation performance for 7 highly ambiguous connectives in the Europarl corpus 67
5.8 Merged data sets for 7 discourse connectives from the Penn Discourse Treebank and the Europarl corpus 68
5.9 10-fold cross-validation performance for 7 connectives with syntactic and semantic features 69
5.10 10-fold cross-validation performance for 7 connectives with feature subset combinations 70
5.11 Significance of feature subset performances for 7 connectives 71
5.12 Disambiguation performance of MaxEnt models for 7 connectives in different test sets 72
5.13 Proportion of labeled connectives in 3 test sets for SMT 73

6.1 Performance of MaxEnt and CRF models on labeling narrativity 76
6.2 Confusion matrix for labeling narrativity 77
6.3 Datasets for English/French verb tense prediciton 80
6.4 Overall performance of MaxEnt models on English/French verb tense prediction 81
6.5 Performance of MaxEnt models for the prediction of specific French verb tenses 82

7.1 English/Czech translation performance with PDTB-labeled connectives with different system combinations and label randomization 88
7.2 Manual evaluation of English/Czech translation performance for discourse connectives with simplified PDTB sense tags 89
7.3 Oracle BLEU performance for an SMT system with verb tense labels 90
7.4 Manual translation evaluation for an SMT system with oracle verb tense labels 91
7.5 Manual evaluation of English/French translation with connectives that were modified in the phrase table to include a sense tag 93
7.6 Automatic and manual evaluation for English/French translation based on SMT systems that were trained on manually and automatically labeled connectives . 95
7.7 BLEU and ACT scores for English/French SMT systems using connective label probability distributions 100
7.8 BLEU and ACT scores for English/French factored phrase-based and hierarchical translation models with POS tags and/or labels for connectives 104
7.9 Genres, sizes and numbers of (labeled) connectives in SMT training, tuning and testing data for 4 language pairs 106
7.10 BLEU and ACT scores of factored translation models with labeled connectives for 4 language pairs 107
7.11 BLEU scores for English/French translation systems based on narrativity labels 110
7.12 Manual evaluation of English/French translations by a narrativity-based SMT system 111
7.13 Manual evaluation of global correctness for English Simple Past verbs translated into French with a system based on narrativity labels 112
7.14 BLEU and classifier performance scores for different SMT system configurations and French tense prediction 113
7.15 BLEU and classifier performance scores for different EN/FR SMT systems per predicted tense 114
7.16 BLEU and classifier performance scores with the best model for French tense predictions 114
7.17 Manual English/French translation evaluation with the best model for French tense prediction 115

8.1 Distribution of PDTB level-2 discourse relations in test data 120
8.2 Counts of implicitation in automatic and human reference translations for English/French and English/German 121
8.3 Distribution of discourse relations in French test data 126
8.4 Distribution of discourse relations in German test data 126
8.5 Counts of explicitation of connectives in automatic and human reference translations for English/French and English/German 127
8.6 Genres, sizes and numbers of connectives in data for building a English/German translation system with implicitation of connectives 132
8.7 System configuration for SMT models with source word deletion features . . . 133

8.8 Manual evaluation of readability for English/German translations with implicitated connectives 137

A.1 Manually compiled list of temporal markers used for verb tense disambiguation 145
A.2 English connectives with a frequency above 20 in the PDTB 147
A.3 French connectives in LexConn 148
A.4 German connectives in DimLex 152

1 Introduction

Machine translation (MT), i.e. the fully automatic translation of text from a natural language to another by means of computer programs, has made tremendous progress in the past two decades. The availability of human-translated, parallel texts (online and elsewhere) as well as the increasing amount of available computing power and memory made it possible to move away from hand-crafted rule-based MT systems to systems that automatically learn statistics and correspondences from large parallel texts in a source and a target language.

MT has reached reasonable performance as long as the source and target language are close in terms of morphology and syntax for instance. In addition, the current statistical MT algorithms only work on a sentence-by-sentence basis and provide accurate translations when considering single sentences independently.

As a consequence, knowledge from previously translated sentences or clauses of a text is lost, as can be illustrated with the example in Figure 1.1, a translation from Google Translate [1], one of the online state-of-the-art MT systems.

English: In terms of the promotion of cultural diversity, which is the more difficult task , **although** I thank you for your efforts at preservation, I am astonished that, ultimately, only audiovisual services **have been retained**.
French-MT: En termes de promotion de la diversité culturelle, qui est la tâche la plus difficile, ***mais** je vous remercie pour vos efforts de conservation, je suis étonné que, finalement, seuls les services audiovisuels ***ont été retenus**.
German-MT: Im Hinblick auf die Förderung der kulturellen Vielfalt, die die schwierigere Aufgabe ist, ***obwohl** ich danke Ihnen für Ihre Bemühungen um Erhaltung, bin ich erstaunt, dass letztlich nur die audiovisuellen Dienste ***wurden beibehalten**.

Figure 1.1: Mistranslations at the discourse level: an example sentence from the Europarl corpus translated from English to French and German using Google Translate.

1. http://translate.google.com/

There are two problems in the French and German translations in Figure 1.1, concerning elementary discourse units (units that establish textual coherence): the discourse connective (*although*) and the verb phrase (*have been retained*). When these are not translated accurately, incoherent translations result in French and German as is the case in the above examples. For the connective *although*, a baseline SMT system cannot grasp that it signals a CONCESSION to what has been previously said. This then leads to an incorrect translation with the FR connective *mais*, here signaling CONTRAST. For German the connective translation with *obwohl* is better as the latter can signal CONCESSION, but not at this syntactic position which moreover greatly decreases the readability of this translation. It would have been more correct to generate the DE connective *jedoch* between *Ihnen* and *für*.

For the translation of the verb phrase *have been retained*, in FR, a specific verb mode is required, because its previous main clause consists of *je suis étonné que* which requires the FR subjunctive mode (the so-called SUBJONCTIF) and should therefore have been translated to *aient été retenus*. In the German translation, the tense of this verb phrase is correct, but the ordering should have been *beibehalten wurden* which again negatively influences the readability of the whole translation.

Discourse connectives and verb tenses are cohesive markers that play an important role for the readability of a text. Connectives relate argumentatively several sentences and signal discourse relations that help the reader to understand causal, temporal or contrastive ordering of clauses and events described. Verb tense, mode and aspect need to be coherently conjugated in a text so that one can correctly deduce the ordering and veridicality of events and states in time. Consequently, the translation of these cohesive markers has to be as appropriate as possible in the target language to convey the source text's exact meaning.

Recent research in MT has tried to address these problems, for example by stacking previously translated units or by two-pass translation strategies that first find and resolve the antecedent of a pronoun and then translate the latter correctly. For lexical consistency there have been attempts to model the topic distribution of a document or to use content word disambiguation to find the most likely senses of the terms used in the document to translate. These approaches can be problematic as they increase the search space (the longer the context the more translation hypotheses have to be generated), or have to rely on (imperfect) pronoun resolution systems. Moreover, they disambiguate content words only, which can help with establishing discourse structure but do not determine it explicitly.

In the present thesis, within the phrase-based statistical MT framework, we have built and evaluated several end-to-end systems that explicitly model two of the above-mentioned cohesive markers that help to establish coherent text structure: connectives and verb tenses. Whereas discourse connectives often only consist of single or multi-word expressions, verb tenses usually are lexicalized as conjugated suffixes on verb stems in the languages studied in this thesis (English, French, German, Italian, Czech and Arabic). We will however show that similar SMT models can be applied to both problems.

The work described in this thesis is among the first to make use of discourse-level features for statistical machine translation (SMT). In the proposed approach, linguistic knowledge is not integrated via hand-crafted rules, but by using classifiers, trained through machine learning, to automatically annotate text-level information – namely, discourse relations expressed by connectives and verb tense labels – in the parallel texts that are needed to train SMT systems. Manual and automatic annotation efforts incorporate features that were found by cross-linguistically analyzing parallel corpora, with the proper disambiguation granularity needed for finding the correct translations. The goal therefore was not only to use the classifiers to improve SMT quality but to thoroughly analyze discourse connectives and verb tenses as translation problems, to find helpful features for disambiguation and to maximize classifier performance. This has indeed a direct influence on the translation output, as the thesis will exemplify. The classifiers perform, in their best configurations, at accuracy levels of 0.7 to 1.0 F1 score and are therefore reliable enough to annotate automatically the large corpora that are used for training SMT systems. Evaluation issues, for disambiguation and for translation, are solved through specific metrics, showing that our SMT system pipelines can improve the translation globally in ranges of 0.2 to 0.5 BLEU points (when using automatic MT scoring) and can improve 2% to 25% of the targeted occurrences of connectives and verb tenses (when evaluating them semi-automatically), depending on the data sets, classifiers and configurations of SMT systems.

In the following sections, we briefly introduce the differences between rule-based and statistical MT approaches and the two collaborative projects in which the author was involved, before providing a detailed overview of the contributions of this thesis.

1.1 Rule-based vs. statistical machine translation

Human manual translation is an expensive and time-consuming activity that needs considerable cognitive and creative effort in order to render and convey a source text's meaning adequately in a target language. Automatic translation by computers, or machine translation (MT), is often referred to as being 'the holy grail' of Artificial Intelligence, Computational Linguistics and Natural Language Processing.

Research in MT has a long and rich history, including decades of enthusiastic exploring as well as ones of disbelief in the field.

After the Second World War, research on this topic began to emerge with attempts to 'decode' Russian texts into English. This terminology that was borrowed from Cryptography is still used today and is also part of a famous quote coined by Warren Weaver in 1949 that would influence the later development of statistical MT:

> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will

now proceed to decode.' [2]

The US military, IBM and other institutions began to encourage MT research through considerable funding and believed that MT would be a solved problem within just a few years. The problem however turned out to be much harder: the divergence between languages in terms of grammar, word order and semantic concepts made it impossible to obtain a coherent output when translating (or decoding) word-by-word only. Therefore, the Automatic Language Processing Advisory Committee (ALPAC) reported in 1966 that more fundamental NLP research was necessary before targeting MT per se, with the direct consequence that funding for MT stopped flowing.

Nevertheless, in the 1970s, a few MT 'pioneers' such as the Systran and Logos companies were established, and the University of Montréal developed an MT system for weather forecast translation. In the 1980s, it was in Japan that the interest for English/Japanese MT systems on personal computers or hand-held devices was highest, while later on, in Europe, the German Verbmobil project at the beginning of the 1990s was successful for speech-to-speech translation. These systems were 'rule-based' ones (RBMT) for which a large set of lexical and/or syntactical transfer rules had to be hand-crafted. This costly procedure also made it hard to adapt these systems to other language pairs, translation directions, or text domains.

The issues of RBMT are often illustrated with the so-called Vauquois-pyramid (Vauquois [1968]) shown in Figure 1.2. In this schema, the system complexity grows when moving to the top, from words only via syntactical transfer rules to a completely language-independent representation (interlingua).

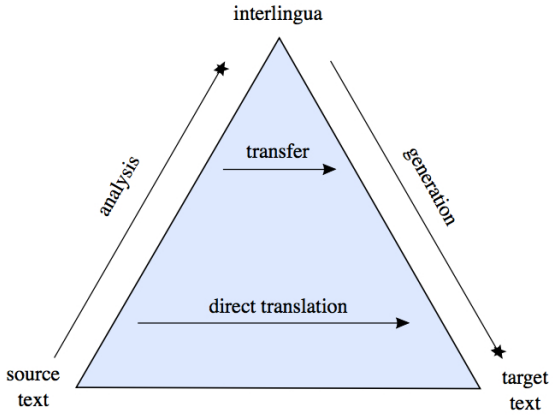


Figure 1.2: The Vauquois pyramid of MT system paradigms and their complexity (Vauquois [1968]). Taken from http://en.wikipedia.org/wiki/Machine_translation.

At the base of the pyramid, or the word level, MT systems perform direct, one-to-one word translations, with possible re-ordering. For most language pairs, this is not sufficient, as

2. Letter by Warren Weaver, March 1947.

either the re-ordering is too distant, or one source word may align to several target words and vice-versa. Deletion and insertion (zero-alignments) pose a further problem at this stage. The MT systems also need to be re-built as soon as a new language pair is added. At the second level of the pyramid, models become more abstract and operate at the syntax level via transfer rules, which are often manually implemented. The advantage is that only a syntax analyzer (on the source side) and a lexical generator (on the target side) are needed, while the transfer rules can be implemented on syntax trees and can possibly be language-independent.

Finally, at the top of the pyramid, a completely language-independent semantic representation of the source text's meaning helps to generate directly the target text. Here, only one analysis and one generation module per language pair need to be implemented. For the semantic modeling of the complete meaning of a text, however, world and domain knowledge is necessary which still cannot extensively be integrated in today's NLP applications. Although there were working systems deploying an interlingua, the idea was abandoned near the end of the 1980's due to exactly this lack of adequate world knowledge models (see for example Nirenburg et al. [2003] and Wilks [2009]).

The past two decades have been ones of new enthusiasm for MT. Throughout the 1990s, SMT was introduced as a promising new paradigm (Brown et al. [1993]). In SMT, where no rule-based processing takes place, the goal is to have a system empirically learn the correct translations of words, phrases and sentences from large collections of human-translated texts, i.e. parallel corpora that have become available in several languages. An often used and cited example of such a corpus is Europarl, containing the statements made by the delegates in the Parliament of the European Union. The corpus provides parallel texts of the statement translations into the 23 official languages of the EU (Koehn [2005]).

In SMT, phrase pairs in source (SL) and target language (TL) can automatically be aligned (at the word level) and corresponding phrase pairs (or chunks of words) can be extracted from a parallel and aligned corpus. To build a translation model, the pairs are accompanied by the number of occurrences, lexical word and phrase translation probabilities, the scores for the reordering of phrases, and their fluency in the TL obtained from a statistical language model. For a more detailed explanation of these translation features, see Section 3.3.

The huge amount of multilingual textual data which has recently become available, together with sophisticated modeling techniques from other related research areas (e.g. speech recognition and machine learning), made possible the training of SMT systems.

Building SMT systems consists of three stages. During the **training** stage, a statistical MT system learns the most likely correspondences and re-ordering of chunks of source words and target words from parallel corpora that have been sentence- and word-aligned. During the **tuning** stage, based on a further (but much smaller) parallel text that ideally is of the same genre as the texts the system should translate in production, the feature weights are optimized for the phrase pairs most likely occurring in this kind of text. The last step is the so-called **decoding** or **testing** stage, at when new texts are translated. During this stage, the

SMT decoder tries to find the most likely phrase pairs from the translation model (phrase table) and re-combines these hypotheses based on probability scores from the translation and language models.

In so-called phrase-based statistical MT systems, the chunks of source and target language words (or phrases) are rather short (normally no more than around 10 words) and are not necessarily linguistically motivated (in terms of sentence constituents, for example). With this procedure and about three million parallel source and target language sentences for training/tuning, over closely related languages, systems can produce translations that are intelligible and, depending on the domain of application, require an acceptable amount of post-editing effort by human translators to reach human-level quality.

One of the most often used, freely accessible and purely statistical MT systems is Google Translate, which is currently able to translate more than 60 language pairs. There even have been attempts to build hybrid, jointly rule-based and statistical MT systems. Systran is the leading example with a free website translator and a commercial system. Apart from these two, there are many other commercial systems for business or personal use, mostly of the hybrid type, such as Language Weaver, Linguatec, or Reverso. During the last decade, also more and more open source decoders and entire SMT toolkits have become available, the most widely used one being the Moses SMT toolkit (Koehn et al. [2007], also see Section 3.3.2).

1.2 Related research projects

The present thesis benefited from the framework of two Swiss SNF Sinergia projects on Machine Translation: COMTIS (CRSI22_127510, 2010-2013) and MODERN (CRSII2_147653, 2013-2016). Both projects have discourse phenomena and SMT as common topic, with slightly differing focus. COMTIS (Improving the Coherence of Machine Translation Output by Modeling Inter-sentential Relations) [3] was a collaboration between the Idiap Research Institute and the Linguistics department and the Computational Linguistics group of the University of Geneva. Empirical, cross-linguistic corpus analyses by the Linguistics department focused on the cohesive markers in question and provided background and features in order to facilitate the manual annotation of the markers, which has mostly been carried out in Geneva as well, with contributions from Idiap, as shown in Chapter 4.

Idiap's main contribution, embodied in this thesis, has been to design and implement automatic classification methods and to integrate the automatic annotation into SMT systems. The Computational Linguistics group at the University of Geneva worked on efficient solutions to speed-up hierarchical translation (e.g. in tree-to-string models), when considering a wider context, and contributed expertise to the integration of text-level features into SMT.

The MODERN project (Modeling Discourse Entities and Relations for Coherent Machine

3. See www.idiap.ch/project/comtis.

Translation) [4] builds upon the work in COMTIS by focusing on lexical consistency in translation at the document level, for noun phrases, pronouns and other means by which reference to entities in discourse is established. This project is a collaboration between the Idiap Research Institute and the Universities of Zurich, Utrecht, and Geneva (Linguistics department). Again, Idiap is responsible for natural language processing and machine learning methods, while the University of Zurich develops MT models incorporating lexical consistency and semantic ontologies. Linguistic issues, such as coherence and readability of translations are studied, e.g. with eye-tracking methods, at the University of Utrecht.

The work described in this thesis is mostly focused on automatic annotation and SMT methodology, although we also present work on manual annotation and on translation evaluation which is directly related to the focus of the thesis, as it enables the construction and evaluation of end-to-end discourse-aware MT systems. This work profited to a large extent from collaboration within the above-mentioned projects. The work described in the last chapter of the thesis has mostly been carried out while the author was at the University of Edinburgh for an internship with Professor Bonnie Webber (February to May, 2013).

At the beginning of each chapter of the thesis and/or in the corresponding sections, we will refer to joint publications and collaborations. When several people contributed to the presented experiments, we will clearly state the contributions of the present author and of the other researchers involved.

1.3 Contributions of the thesis

This thesis makes several contributions to the field, from data generation and annotation to complete, end-to-end discourse-aware SMT systems, which improve the translation of the targeted word types. We present here an overview of the content and contributions of the thesis, which will be discussed in detail in the corresponding chapters.

Chapter 2. Discourse connectives and verb tense in translation

This chapter introduces the translation problems that are related to discourse connectives and verb tenses. We will exemplify, with human reference translations and output by current state-of-the-art SMT systems, several problematic cases, i.e. when a discourse connective needs to be disambiguated prior to translation because the target language does not preserve the ambiguity of the source language's marker or translates the source language marker by other lexical and syntactical means or not at all. Similarly, the verb tense systems of the source and the target languages can diverge significantly, and for a source language tense, several possible target language tenses are available, among which the correct one has to be found depending on the longer-range context of the current discourse. We will return to similar examples when modeling the two discourse phenomena later on in the corresponding chapters.

4. See www.idiap.ch/project/modern.

Chapter 3. Related work

This chapter presents previous work related to the various contributions of the thesis and discusses from this perspective the novelty of our proposals.

Chapter 4. Data, annotation procedures and evaluation metrics

Over the recent years, more and more discourse-annotated resources have become available. However, they are most often monolingual only and no human reference translation exists, against which an SMT system could be evaluated.

After thorough theoretical and empirical analyses of discourse relations and verb tenses, and of the related translation divergencies, manual gold-standards of several thousand sentences have been created by labeling occurrences in the above-mentioned Europarl corpus, with trained annotators. We carefully extracted texts that have been recorded in original source language and their corresponding direct translations, that are not distorted by an already translated source or an intermediate pivot language. Where the manual annotation was difficult, which was the case for most of the ambiguous connectives dealt with, a method called *translation spotting* was used to generate reliable gold-standards. Theoretical analyses and development of the annotation guidelines and procedures have been carried out together with the Linguistics Department of the University of Geneva.

For verb tenses, the approaches to manual annotation were slightly different. For one series of experiments, two trained annotators identified manually in a corpus of different genres whether an English Simple Past verb was used in a narrative or non-narrative context. For finding and disambiguating more English tenses for translation into French, an automatic method for the generation of a large training set was used. The method extracts aligned verb forms in English and French based on word alignment, dependency parsing and French morphological analysis. These resources for verb tenses were again produced in close collaboration with the Linguistics Department of the University of Geneva.

The resources are freely available to the community for further research and replicability [5]. Moreover, Chapter 4 presents the evaluation metrics used in this thesis.

Chapter 5. Automatically disambiguating discourse connectives

When humans process a coherent text, it has been shown that the correct usage and placement of connectives influences the efficiency and adequacy of inference of the argumentative structure and meaning of the text. Wrongly generated connectives (e.g. produced by an SMT system) therefore affect the coherence and in consequence the quality of the text perceived by its reader.

5. See www.idiap.ch/dataset/disco-annotation for discourse connectives and www.idiap.ch/dataset/tense-annotation for verb phrases.

For an automatic discourse processing component, the detection of discourse connectives and the discourse relations they signal is an important step as textual coherence or discourse structure is often established by connectives that relate spans of text and indicate information about temporal ordering, causality and/or contrast.

Discourse connectives can *ambiguously* signal these relations, depending on the set of such markers available in a language. Human readers can most often reliably determine the correct meaning of a connective from the available context or from world knowledge. Neither world knowledge nor structural inference are normally available to NLP systems, and the features that can be extracted from a connective's context might not be sufficient to point to the signaled relation. The disambiguation of connectives might, at first hand, look like a word sense disambiguation (WSD) problem, where the same words (e.g. *bank*) can have different meanings in different contexts (*river bank* vs. *money bank*). For WSD however, features from the close context are most often sufficient to find these meanings. As discourse connectives can relate separate sentences, sometimes even within different paragraphs, finding the signaled relation can be difficult for humans and especially for NLP systems. The automatic disambiguation of connectives therefore is an open research problem.

For some of the most ambiguous connectives, we introduce a number of new and helpful features that help to learn automatically the relations that they signal. The set of relations we use is sometimes more detailed as the ones used in state-of-the-art systems. In the latter, often only the top classes of a discourse relation taxonomy are used, whereas we try to classify relations at a more detailed level and we also account for instances where a connective may signal two discourse relations at the same time. The developed feature extractor and the trained disambiguation models are freely accessible for other researchers in order to annotate new texts and for comparison of results [6].

An analysis of possible translation errors regarding connectives is presented in Chapter 2, while Chapter 5 is dedicated to the automatic disambiguation of discourse connectives.

Chapter 6. Automatically disambiguating verb tense

Similar to the importance of connectives for human and automatic text processing, the correct usage of verb tense influences textual coherence in terms of relating the events and states described in a text into the correct temporal ordering. Translating to a wrong verb tense can go as far as misleading the reader in terms of whether an event actually happened or when it happened within the overall narrative of the text.

Few previous studies exist on verb tense disambiguation and translation. Via contrastive linguistic analyses, we however identified the most frequent translation divergencies for the English/French language pair, where for example the English Simple Past tense poses the most problems as there is no one-to-one mapping from its English usages to the French ones:

6. www.idiap.ch/dataset/disco-annotation

at least three different tense forms are indeed valid translations depending on the narrative context.

We implemented two disambiguation systems, one for a binary discursive feature called *narrativity*, the other as a French tense predictor that automatically outputs the most likely French tense an English verb should be translated to. As with discourse connectives, these models and their feature extractors are publicly available for further research [7].

Moreover, Chapter 2 provides an error analysis for verbs in current statistical MT systems.

Chapter 7. Statistical machine translation with discourse labels

For statistical MT that inherently does not make use of any linguistic information or rules, it is not obvious how to model phenomena that take place beyond the sentence level.

In this work, we implemented system pipelines that can directly make use of the linguistic information the automatic disambiguation modules have annotated in the training and testing data sets. This can be seen as a pre-processing step that modifies the raw text of the translation input so that the discourse information is present in forms of labels on connectives and verb forms.

Frameworks for such a pre-processing step have partially already been available for so-called factored and/or hierarchical syntactic SMT, where either morphological or tree-like grammar structures are integrated in the SMT training procedure. We compare against a number of own approaches, such as system combination (baseline and discourse-aware systems) or label/data distribution based on classifier confidence for its predictions.

As the tools we need for feature extraction are to a vast extent only available when processing English, most of our experiments are for systems that translate from English to another language: French, German, Italian, Arabic and Czech – thus illustrating the generalizability of our methods.

For all types of experiments we provide thorough analyses of the output, and translation quality evaluation. Current automatic MT metrics such as the BLEU score are not yet sensitive enough to capture or account for the few word changes our models perform. We therefore resorted to semi-automatic measurements, where we compare a system's output translation versus its human reference and/or a baseline system which did not involve any discourse-level features. The improvements in terms of connectives and verb tenses are then counted as the variation of the percentage of items that are better, equally or less well translated by our system compared to a baseline system but also to reference translations.

Another approach to evaluation is to automatically count correct translations of discourse connectives based on word alignments of a system's output with the source text and by

7. www.idiap.ch/dataset/tense-annotation

comparing to a human reference. To achieve maximum precision, this metric relies on dictionaries of discourse connectives and their valid translations, including synonyms (for details, see Chapter 4).

All in all, automatic MT scores such as BLEU show that our modified, discourse-aware models do not significantly degrade scores but rather improve them, by 0.2 to 0.5 BLEU points, especially for verb tenses, which are much more numerous than connectives. Manual and semi-automatic scores confirm that our systems translate the targeted discourse units more correctly than the counterpart baseline systems, in ranges of 2% to 25%, depending on the performance of the automatic classifiers, the data sets used and the system configurations implemented.

Chapter 8. Statistical machine translation with deletion/insertion of connectives

As a final contribution we had a close look at the situations in which humans tend to omit a discourse connective in the target language where there has been one in the source (*implicitation*) or, vice-versa, to the situations in which they introduce a target language connective where there was no such word in the source language (*explicitation*). Could this be modeled accurately in SMT systems, their output would be made more similar to human translation and hence more fluent and more natural, besides being more similar to the reference translation.

Analyses on parallel texts revealed the high frequency with which human translators perform such insertions and deletions and how much less often this is the case for current SMT systems, which remain closer to the wording of the source text. We have undertaken first steps via new features in SMT decoding to account for the implicitation of connectives in automatic translation.

These experiments are presented in Chapter 8 of the thesis, before concluding with perspectives on future work in Chapter 9.

2 Discourse connectives and verb tense in translation

In this chapter, we introduce two cohesive markers, discourse connectives and verb tense, and present the cases of mistranslation that occur for them with current SMT systems. The goal of the thesis is to avoid translation errors for these two cohesive markers as their mistranslation can lead to incoherent, distorted target language text which is contrary to the goal of correct MT that should be driven toward coherent translation output.

Translation errors for connectives and verb tense, as it will be illustrated with concrete translation examples, can be as severe as misleading the reader with MT output that might grammatically be correct, but does not reflect the same meaning as the source text had. In less severe cases the reader still can infer, with context and word knowledge, what the meaning should have been.

Although there have been attempts in previous research to address lexical consistency throughout entire documents instead of sentences only, these focused on content words most of the times and are often word sense disambiguation methods coupled with SMT (reviewed in Chapter 3, Section 3.3.2).

Instead, we here focus on a specific type of function words, discourse connectives, that have a procedural role in linking spans of texts or even paragraphs in a meaningful way (Section 2.1). The correct marking of verb tense similarly ensures that the information described in a text appears in a meaningful order related to its appearance in the real world (Section 2.2).

Cohesive markers refer to the linguistic devices that establish coherence between spans of text. In early research on such linguistic devices, cohesive markers have already been considered to be lexical and grammatical items like pronouns and referential expressions, discourse connectives and verbal tenses (Halliday and Hasan [1976]). Coherence is a universal property of discourse, whereas each language varies in terms of the set of cohesion markers available even when they are closely related, such as English and French.

2.1 Discourse connectives in translation

Discourse connectives are a class of frequent cohesive markers, such as *although, however, for example, in addition, since, while, yet*, etc. They are function words of rather low frequency compared to other words in a text. For instance, in the Penn Discourse Treebank (Prasad et al. [2008]) (see Section 4), 1.8% of the 1'000'000 tokens from the Wall Street Journal corpus are annotated as discourse connectives. The actual set of markers or connectives is however rather open-ended (Prasad et al. [2010]), as there are multi-word expression such as *at the same time, over all, given that* etc. that can serve as a connective. Connectives contribute to the establishment of argumentative textual structure and high-level understanding of relations between sentences.

A discourse connective can, in some contexts be ambiguous and for example either convey temporal ordering or causal relationship between the two text spans linked.

Moreover, the same connective can simultaneously convey more than one discourse relation. For example, *while* can convey contrast or a temporal meaning (simultaneity), or both at the same time. On the other hand, discourse relations can also be conveyed implicitly, without an explicit connective.

2.1.1 Translation problems related to connectives

For many occurrences of English connectives, determining the exact relation that they signal is necessary for correct translation because the target language may have a different set of connectives available and/or those available may not be of the same ambiguity as the source language connectives. However, most current SMT models use features that are too local to allow modeling the ambiguities of discourse connectives. Therefore, the translation of ambiguous connectives is often mistaken, which has a detrimental impact on the coherence and readability of SMT output.

Connectives are furthermore especially prone to 'translationese', i.e. the use of constructions in the target language that differ in frequency or position from how they would be found in texts originally written in that language. They can be translated in ways that can differ markedly from their use in the source language. For cohesive markers and discourse connectives, Koppel and Ordan [2011], Cartoni et al. [2011], Ilisei et al. [2010] and Baroni and Bernardini [2005] have shown that there may be more explicit (increased use) or less explicit (decreased use) in translationese. Translated language can be simpler (lexically less dense) (Laviosa-Braithwaite [1996]) and consisting of fewer items that are unique to the target system (i.e. items without exact equivalents in the source language) (Tirkkonen-Condit [2002]).

Human translators can choose to not translate a source language connective with a target language connective, where the latter would be redundant or where the source language discourse relation would more naturally be conveyed in the target language by other means

(cf. Chapter 8).

We will use the term 'zero-translation' or 'implicitation' for a valid translation that conveys the same sense as a lexically explicit source language connective, but not with the same form. The latter can be more natural in many cases, but for MT to simply delete or not translating a connective regardless of its context can lead to incoherent target text inasmuch as a wrong connective would do. As we will show, current SMT models either learn the explicit lexicalization of a source language connective to a target language connective, or treat the former as a random variation, realizing a connective word form or not.

Learning the correct target connective and/or other valid ways of conveying the same discourse relation would not only result in more fluent target language text, but also help raise automated MT evaluation scores because a system output would be more closely resembling its human reference text.

2.1.2 Examples of translation errors

The following examples illustrate these types of errors. In EN, the discourse connective *since* can signal two principal discourse relations: TEMPORAL and CAUSAL and in rare cases both at the same time. In contrast, in FR, there are different lexical connectives for these relations: *depuis (que)* for TEMPORAL and *parce que, car, puisque* for CAUSAL.

In Example 2.1, for the human reader, *since* in EN quite clearly signals a temporal relation, although the syntax of the sentence could as well introduce a reason for why the doctrine has been criticized. In the human FR reference translation, *since* is correctly translated to *depuis* with a temporal meaning. A baseline SMT system however, due to phrasal constraints, generated the connective *parce que*, unambiguously signaling a causal discourse relation which in this context leads to a possible interpretation of the target text which is however different from the intended one: the doctrine has been criticized '*because* it was published first' and not '*during* the time of its publication', the latter being the EN source text's original meaning.

English: What stands between them and a verdict is this doctrine that has been criticized **since**_TEMPORAL it was first issued.
French-Reference: Seule cette doctrine critiquée **depuis**_TEMPORAL son introduction se trouve entre eux et un verdict.
French-Baseline-MT: Ce qui se situe entre eux et un verdict est cette doctrine qui a été critiqué ***parce qu'**_CAUSAL il a d'abord été publié.

Figure 2.1: Mistranslation of a discourse connective from English (*since*) to French (reference: *depuis*, MT: **parce que*). The example comes from the nt2012 data set described in Section 7.7.2.

Example 2.2 illustrates a translation from EN to DE, where not translating a discourse connec-

tive at all has the same consequence as for the EN/FR example: the target text is well-formed but has the very opposite meaning. The EN connective *while* signals a CONTRAST relation, which is entirely missing from the baseline MT translation in DE. As a consequence, the sentence reads more like: 'I welcome the rapid action *and* we have to be clear' instead of 'I welcome the rapid action *but* we have to be clear'. In the human reference translation, the latter is correctly rendered by the DE contrastive connective *zwar* and is reinforced by a second explicitated contrastive marker: *doch*.

English: Thirdly, **while**_CONTRAST I welcome the rapid reaction force, we have to be clear from Europe's perspective, as we only get one chance to get this right.
German-Reference: Drittens halte ich **zwar**_CONTRAST die schnelle Eingreiftruppe für begrüssenswert, doch müssen wir eindeutig aus europäischer Sicht handeln, denn uns steht nur eine Chance zur Verfügung, es richtig zu machen.
German-Baseline-MT: Drittens, ***__0__** ich begrüsse die schnelle Eingreiftruppe, müssen wir uns im Klaren sein in der europäischen Perspektive, wie wir nur noch eine Chance, dieses Recht.

Figure 2.2: The discourse connective *while* is not translated at all from English to German in MT, leading to the opposite meaning that is established in the human reference by the DE connective *zwar* (from nt2008, see Section 7.7.2).

Besides these misleading cases, there are less severe contexts, in which a discourse connective is interchangeable with another one without losing meaning and/or grammaticality. In Example 2.3, a translation from EN to IT, the human translator chose to use the connective *sebbene* for the EN *even though*, signaling CONCESSION. The EN/IT MT baseline system generates an explicit discourse connective, *anche se*, which is in this context correct and equivalent in terms of position and signaled discourse relation.

English: Mr President, we are debating the third agreement with Morocco, which above all concerns French and Portuguese fishermen, who make up the bulk of the fleet, **even though**_CONCESSION a small number of French and Swedish fishermen are also concerned.
Italian-Reference: Signor Presidente, ci troviamo dinanzi al terzo accordo con il Marocco, che interessa soprattutto gli Spagnoli e i Portoghesi, ovvero il grosso di la flotta, **sebbene**_CONCESSION siano ugualmente interessati anche un discreto numero di Francesi e di Svedesi.
Italian-Baseline-MT: Signor Presidente, stiamo discutendo di la terza accordo con il Marocco, che riguarda soprattutto i pescatori Francesi e Portoghesi, che costituiscono la maggior parte di la flotta, **anche se**_CONCESSION un piccolo numero di, i pescatori Francesi e Svedesi sono anche preoccupato.

Figure 2.3: Example of equivalent IT discourse connectives (*sebbene* and *anche se*) in the human reference translation and the MT output for the EN connective *even though* (from nt2008, see Section 7.7.2).

In all the cases described above, a label indicating the discourse relation signaled by a connective would be sufficient in order to find its correct target language equivalent or a synonym thereof. Annotating that information prior to translation and making an SMT system learn from it so that it improves its output is the hypothesis our thesis work started from. In the following, we illustrate that the same idea holds true as well for the translation of another cohesive marker, i.e. verb tense.

2.2 Verb tense in translation

The text in Figure 2.4 is an example of a four-sentence discourse, in which the English verbs, all in Simple Past tense (SP), express a series of events having occurred in the past, which no longer affect the present. As shown in the French translation by a baseline SMT system (not specifically aware of verb tense), the English SP verbs are translated into the most frequent tense in French, as learned from the parallel data the SMT was trained on.

When looking at the example more closely, however, it appears that the SP actually conveys different temporal and aspectual information. The verbs *offered* and *found* describe actual events that were ordered in time and took place in sequence (hence a narrative context), whereas *were* and *was* describe states of general nature, not indicating any temporal ordering (hence a non-narrative context).

EN: (1) After a party, I offered [**Narrative**] to throw out a few glass and plastic bottles. (2) But, on Kounicova Ulice, there were [**Non-narrative**] no colored bins to be seen. (3) Luckily, on the way to the tram, I found [**Narrative**] the right place. (4) But it was [**Non-narrative**] overflowing with garbage.

FR from BASELINE MT system: (1) Après un parti, j'**ai proposé** pour rejeter un peu de verre et les bouteilles en plastique. (2) Mais, sur Kounicova Ulice, il n'y **avait** pas de colored bins à voir. (3) Heureusement, sur la manière de le tramway, j'**ai trouvé** la bonne place. (4) Mais il ***a été** débordés avec des ramasseurs.

Figure 2.4: Example English text from the 'nt2010' data with narrativity labels and a translation into French from a baseline SMT system. The tenses generated in French are, respectively: (1) Passé Composé, (2) Imparfait, (3) Passé Composé, (4) Passé Composé. The mistake on the fourth one is explained in the text.

The difference between narrative and non-narrative uses of the EN SP is not always captured correctly by the baseline SMT output in this example. The verbs in the first and third sentences are correctly translated into the French Passé Composé (PC) (one of the two tenses for past narratives in French along with the Passé Simple (PS)). The verb in the second sentence is also correctly rendered as Imparfait (IMP), in a non-narrative use. However, the verb *was* in the fourth sentence should also have been translated as an IMP, but from lack of sufficient information, it was incorrectly translated as a PC (moreover, with the wrong mode and past

participle agreement). A non-narrative label could have helped to find the correct verb tense, if it would have been annotated prior to translation.

The difficulty for MT systems is thus to choose correctly among the three above-mentioned tenses in French, which are all valid possibilities for translating the English SP. When MT systems fail to generate the correct tense in French, several levels of incorrectness may occur, similar to the situation of connectives. These levels are exemplified in Figure 2.5 with sentences taken from the data by Grisot and Cartoni [2012].

1. In certain contexts, tenses are interchangeable, which is the unproblematic case for MT (although single-reference evaluation metrics will penalize a variation). In Example 1 from Figure 2.5, the verb *étaient considérées* (were seen) in IMP has a focus on temporal length which is preserved even if the translated tense is a PC (*ont été considérées*, i.e. have been seen) thanks to the adverb *toujours* (always).
2. In other contexts, the tense proposed by the MT system can sound strange but remains acceptable. For instance, in Example 2, there is a focus on temporal length with the IMP translation (*voyait*, viewed) but this meaning is not preserved if a PC is used (*a vu*, has viewed) though it can be recovered by the reader.
3. The tense output by an MT system may be grammatically wrong. In Example 3, the PC *a renouvelé* (has renewed) cannot replace the IMP *renouvelaient* (renewed) because of the conflict with the imperfective meaning conveyed by the adverbial *sans cesse* (again and again).
4. Finally, a wrong tense in the MT output can be misleading, if it does not convey the meaning of the source text but remains unnoticed by the reader. In Example 4, using the PC *a été* leads to the interpretation that the person was no longer involved when he died, whereas using IMP *était* implies that he was still involved, which may trigger very different expectations in the mind of the reader (e.g. on the possible cause of the death, or its importance for the peace process).

Instead of annotating tense information via the binary label of narrativity (i.e. narrative or not), with which only the EN SP can be processed, another possibility is to try to predict automatically the FR tense that should be used for each EN verb in context. By analyzing translation data from a large parallel corpus, one can count the frequency and distribution of EN and FR verb tenses and their (non-)correspondences (Table 2.1). For 322'086 verb phrases, we found these percentages of verb tense translations, illustrating the largest EN/FR divergencies. For the method used to create these counts, see Section 4.2.2 and the collaborative paper (Loaiciga et al. [2014]).

In Example 2.6, there are 5 EN verbs in the sentences: 3 in Present tense, 1 in Future tense, and 1 in Simple Past. In the FR human reference translation, one can see that these EN tenses

1. EN: Although the US viewed Musharraf as an agent of change, he has never achieved domestic political legitimacy, and his policies **were seen** as rife with contradictions.
FR: Si les Etats-Unis voient Moucharraf comme un agent de changement, ce dernier n'est jamais parvenu à avoir une légitimité dans son propre pays, où ses politiques ont toujours **été considérées** (PC) / **étaient considérées** (IMP) comme un tissu de contradictions.

2. EN: Indeed, she even persuaded other important political leaders to participate in the planned January 8 election, which she **viewed** as an opportunity to challenge religious extremist forces in the public square.
FR: Benazir Bhutto a même convaincu d'autres dirigeants de participer aux élections prévues le 8 janvier, qu'elle **voyait** (IMP) / **?a vu** (PC) comme une occasion de s'opposer aux extrémistes religieux sur la place publique.

3. EN: The agony of grief which overpowered them at first, **was** voluntarily **renewed**, was sought for, was created again and again...
FR: Elles s'encouragèrent l'une l'autre dans leur affliction, la **renouvelaient** (IMP) / l'***a renouvelé** (PC) volontairement, et sans cesse...

4. EN: Last week a person who **was** at the heart of the peace process passed away.
FR: La semaine passée une personne qui **était** (IMP) / **#a été** (PC) au cœur du processus de paix est décédée.

Figure 2.5: Examples of translations of the English SP by human translators and a baseline SMT system, differing from the reference translation: (1) unproblematic, (2) strange but acceptable (?), (3) grammatically wrong (*), and (4) misleading (#).

do *not* have, in this context, direct correspondences and the FR verbs are conjugated by two present tenses, one in future tense and one verb (to object) is actually not translated as verb but noun phrase (l'*objection*).

In order to render the same FR tense information needed to translate the EN verbs correctly, an idea is to annotate, onto the EN verbs, a label that directly consists of the FR tense, if it can be predicted automatically with sufficient accuracy.

After a review of the related work in the following chapter, Chapter 4 will provide an overview of the language resources we made use of in order to label discourse relations and verb tenses, either manually, as will be described in the same chapter, or automatically, as will be the topic of the chapters following it (Chapters 5 and 6).

EN: Madam President, if the vote **records_PRES** correctly how my Group **voted_SIM_PAST** I **shall_FUT** not, and **cannot_PRES**, **object_PRES** to that.
FR-Ref: Madame la Présidente, si le procès-verbal **reflète_PRÉS** correctement le vote de mon groupe, je n'**ai_PRÉS** et n'**aurai_FUT** aucune objection à formuler.
EN-MT-input: Madam President, if the vote **records_PRÉS** correctly how my Group **voted__0__** I **shall_PRÉS** not, and **cannot_FUT**, **object__0__** to that.

Figure 2.6: Example of verb divergencies between an English source text and its French reference translation. The third item shows a possible input to an MT system, where the EN verbs have been labeled with information from the FR reference tenses.

FR/EN	P_cont	P_perf_c	P_perf	PRE_cont	PRE_perf_c	PRE_perf	PRE	P_simp	Total
Imparf.	53.5%	**26.9%**	**24.4%**	0.8%	1.8%	1.1%	0.7%	**20.5%**	3.4%
Impér.	–	–	–	0.2%	0.1%	0.0%	0.1%	0.0%	0.1%
Passé Comp.	16.1%	7.7%	14.3%	1.5%	**33.3%**	61.3%	0.6%	49.3%	15.0%
Passé Réc.	–	–	0.1%	0.0%	0.3%	0.4%	0.0%	0.0%	0.1%
Passé Simp.	0.5%	–	0.4%	0.1%	0.2%	0.1%	0.0%	1.0%	0.2%
P.-q.-parf.	3.1%	**30.8%**	52.2%	0.0%	0.4%	0.5%	0.0%	2.9%	0.7%
Prés.	**25.0%**	**34.6%**	6.8%	96.0%	63.2%	**34.1%**	97.2%	**24.9%**	79.1%
Subj.	1.7%	–	1.9%	1.4%	0.6%	2.4%	1.4%	1.4%	1.5%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 2.1: Distribution of EN/FR verb tense translations in the Europarl corpus, over 322'086 verb phrases. The abbreviated tenses are the following: for EN: P_cont = Past continuous, P_perf_c = Past perfect continuous, P_perf = Past perfect, PRE_cont = Present continuous, PRE_perf_c = Present perfect continuous, PRE = Present tense, P_simp = Past simple; for FR: Imparf = Imparfait, Impér. = Impératif, Passé Comp. = Passé Composé, Passé Réc. = Passé Récent, Passé Simp. = Passé Simple, P.-q.-parf. = Plus-que-parfait, Prés = Présent, Subj = Subjonctif. The most prominent translation divergencies are highlighted in bold.

3 Related work

3.1 Discourse processing

The disambiguation of discourse connectives is a task related to discourse parsing. In the latter, however, entire discourse structure trees are inferred automatically, whereas disambiguation of (explicit) connectives can be achieved by locating their word form and by deriving (sometimes complex) features from their context in order to find the discourse relation they signal.

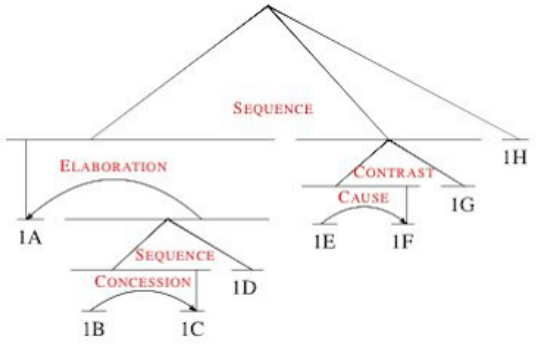
3.1.1 Discourse parsing

Besides morphological, syntactical and semantic analysis in NLP, analysis at the discourse level has long been recognized as useful and necessary in order to deal with entire paragraphs and documents that do not contain phrases and sentences in isolation, but consist of a coherent textual structure that reflects the author's intention. An author of a text usually arranges text segments in a temporally, causally or argumentatively meaningful order.

Discourse processing therefore can start as early as finding those textual segments or so-called 'elementary discourse units' (EDUs) (Marcu [2000]) that provide in itself information but at different importance levels. Discourse relations between EDUs, such as CAUSE, CONTRAST, ELABORATION etc., help the reader to infer the ordering of the information and events described.

In discourse representation theories, e.g. Rhetorical Structure Theory (Mann and Thompson [1988]), a text is often represented as (binary or greater) discourse tree, similar to the syntactical tree structures of generative grammar, for example. Instead of sentences and their constituents, the discourse tree links paragraphs and spans of text or EDUs. The leaves of the tree can either be nuclei (EDUs that provide the minimal information to understand the text), or satellites (EDUs that provide additional information), both linked, at the branches, by discourse relations that establish the tree structure. The links themselves are sometimes referred to as being 'paratactic' (for links between nuclei only) and 'hypotactic' (for links of the

type nucleus-satellite).



[Yesterday, the delegates chose their new representative.]^{1A} [Even though Smith received only 24 votes,]^{1B} [he accepted the election with a short speech.]^{1C} [Then the assembly applauded for three minutes.]^{1D} [Due to the upcoming caucus meeting,]^{1E} [the subsequent discussion was very short.]^{1F} [Nonetheless the most pressing questions could be resolved.]^{1G} [The meeting was closed at 7pm.]^{1H}

Figure 3.1: Example RST tree with four paratactic discourse relations (SEQUENCE) and four hypotactic ones (ELABORATION, CONCESSION, CONTRAST, CAUSE). Figure taken from http://www.david-reitter.com/compling/rst/index.html

As can be seen from the example tree in Figure 3.1.1, the linking discourse relation is often lexically signaled by a discourse connective, placed for instance at the beginning of satellite EDUs (see *even though*, *then*, *due to* and *nonetheless* in the sentences above). The last sentence is an example of an implicit temporal discourse relation that can be inferred by the reader as being a continuation of the top SEQUENCE relation.

In approaches to discourse processing (Marcu [2000], Soricut and Marcu [2003], Le Thanh et al. [2004], Lüngen et al. [2006]), the next step after having found the EDUs, is to infer from the EDUs the entire tree structure, by concatenating identified EDUs following a set of rules that mostly rely on the cue phrases present and punctuation in their context. Marcu et al. [2000] (see also Section 3.3) have proposed an RST-based model for the translation of discourse structure from Japanese into English, but no MT results were reported, which is why our work is among the very first to integrate discourse structure into fully functional MT systems.

Recent discourse parsers try to learn the discourse structure automatically from a large amount of mostly hand-labeled data and rely, instead of hand-made tree-building rules, on machine learning algorithms such as support vector machines, maximum entropy algorithms or structural learning (see e.g. Wellner [2009], Hernault et al. [2011], Lin et al. [2014]). Discourse parsing has proven to be a difficult task, even when complex models are used. The performance of discourse parsers is in a range of 0.4 to 0.6 F1 score [1]. Lin et al. [2014] recently released a discourse parser that labels rhetorical relations and the linked text spans in PDTB style.

Discourse parsing remains an unsolved problem for several reasons, one being already at the very first step in the processing pipeline: the manual annotation of complex RST or other

1. When calculating performance for discourse parsing, *precision* is the percentage of discourse relations with a specific type in the parser output that were correct, *recall* is the percentage of discourse relation with a specific type in the test set that were correctly parsed and the *F1 score* is the harmonic mean of precision and recall.

theory-based discourse trees is a difficult task. Abstracting from the paragraph to the entire text level leads to more subjective choices and a higher rate of inter-annotator disagreement. When these cases are resolved to one of the annotators' decision or even discarded, it might lead on the one hand to more tractable and machine learnable resources, on the other hand, however, this will not help in advancing the theory and hence automatic discourse processing will not advance in dealing with difficult tree structures. Corpora where ambiguity and cases of doubt are preserved would therefore be useful in future work to study the (automatic) building of complex discourse structures (Stede [2011]).

All the above has lead to establish the (more tractable) disambiguation of the connectives and finding the relations they signal as a task in its own right. As our goal here is, chiefly, to study the applicability of discourse-level features to SMT in order to translate connectives more correctly, we follow this approach and have implemented classifiers with an extensive feature set for the connective disambiguation task.

3.1.2 Disambiguating discourse connectives

One of the first studies on identifying discourse connectives and the relations they signal suggested that most English connective types are rather easy to identify, as they occur in unambiguous usages (Pitler et al. [2008]). The state-of-the-art performance for labeling all types of connectives in English is therefore quite high. When using the Penn Discourse Treebank (PDTB) (Prasad et al. [2008]), as training, development and test data, for example, the disambiguation of discourse vs. non-discourse uses of connectives reaches 97% accuracy (Lin et al. [2014]) [2]. The labeling of the four main senses from the PDTB sense hierarchy (temporal, contingency, comparison, expansion) reaches 94% accuracy (Pitler and Nenkova [2009]) – however, the baseline accuracy is already around 85% when using only the connective token as a feature. Various methods for classification and feature analysis have been proposed: Wellner et al. [2006], identification of a connective's argument spans; Wellner and Pustejovsky [2007], usefulness of features for temporal ordering of events; Elwell and Baldridge [2008], argument identification with connective-specific classifiers.

This picture drastically changes when one tries to disambiguate only certain, highly ambiguous types of connectives or ones that pose problems in translation, as we do here. Only a few studies have focused on the analysis of highly ambiguous discourse connectives only. Miltsakaki et al. [2005] report classification results for the connectives *since*, *while* and *when*. Using a maximum entropy classifier, they reach 75.5% accuracy for *since*, 71.8% for *while* and 61.6% for *when*. As the PDTB was not completed at that time, the data sets and labels are not exactly identical to the ones that we will use in this thesis.

Versley [2011] designed hierarchical maximum entropy classifiers for the PDTB hierarchy of labels, going down to the third sense level, and using syntactical and verbal tense/mood fea-

2. The PDTB is one of the largest hand-annotated resource for discourse connectives, discourse relations and the text spans they are linking. Please see Chapter 4 for a more detailed description of the corpus.

tures. The author provides detailed results for up to 25 single connectives, with performances in a range of 45% to 100% accuracy, with the most difficult distinctions being CONTRAST vs. CONCESSION and TEMPORAL vs. CONTINGENCY. The studies by Miltsakaki et al. and Versley are in line with ours and confirm the increased difficulty when (a) disambiguating single, highly ambiguous connectives only and (b) this disambiguation aims for detailed PDTB senses of the second and third PDTB hierarchy levels. We will compare our proposal more closely to these two studies when reporting on our own disambiguation experiments, in Chapter 5.

Obtaining better results when classifying for specific types of discourse markers with a single classifier for each, instead of classifying all types jointly, has also been demonstrated by Popescu-Belis and Zufferey [2007], in the case of discourse markers *like* and *well*. These markers were found to be more accurately identified when processed separately. As with discourse connectives (that sometimes are regarded as discourse markers as well), there does not seem to be a notion of a homogeneous class and many features to either find discourse usage or the relations signaled are item-specific, as was found by Litman [1994] already. Although we will compare our discourse connective type-specific classifiers to joint ones, we can already stress here that we always used item-specific classifiers to annotate the training data for the SMT systems, in order to reach the most reliable automatic annotation before translation.

In all of the above-mentioned work, it has been shown that features at the syntactical level – such as the constituent path leading to the connective and the categories of the words present in its context – account for most of the performance for discourse relation disambiguation. Given that we would like to classify a specific subset of highly ambiguous connectives that are problematic for translation, we have also implemented a series of more semantically-oriented features and will compare their usefulness against the state-of-the-art in Section 5.6.2.

3.2 Modeling verb tense

Verbs are essential to language because they declare states and actions. Moreover, verbs convey various indications of tense, aspect and mode (TAM). In other words, not only do they declare that an event takes place, but place it in a particular time, encode the perception of the speaker about it, and express the level of factuality (Aarts [2011]). These categories, however, interact and overlap, and are used differently across languages.

For instance, when translating verbal phrases (VPs) into a morphologically rich language from a less rich one, mismatches of the TAM categories arise. The difficulties of generating highly inflected Romance VPs from English ones have been noted for languages such as Spanish (Vilar et al. [2006]) and Brazilian Portuguese (Silva [2010]). Samardzic et al. [2010] have studied translation divergencies regarding predicate-argument structures (semantic roles such as subject (agent), object (patient), theme, experiencer etc.) for English/French, finding only about 5% mismatching predicate-argument structures which indicates that a great majority of French predicates directly corresponds to an English verb with the same

predicate-argument structure [3]. For languages pairs less related this rate can be much higher, i.e. 30% for English/German or 17% for English-Chinese.

In the present thesis we mostly focus on tense and aspectual information, for which, as we have already shown above (Section 2.2), considerable translation divergencies exist, in particular for the problem of translating EN past tense to FR ones. Although there are a number of existing and annotated resources for the automatic processing of verbs and VPs (TimeML http://timeml.org/site/index.html, FrameNet (Baker et al. [1998]), VerbNet (Kipper [2005]) or PropBank (Palmer et al. [2005]), they all come, similarly to the PDTB, with the disadvantage of being monolingual (English) only and therefore do not offer the coverage needed for unrestricted MT.

Regarding the specific translation divergency for EN/FR in terms of past tense, the classical view on verb tenses that express past tense in French (Passé Composé (PC), Passé Simple (PS) and Imparfait (IMP)) is that the PC and PS are both perfective, indicating that the event they refer to is completed and finished (Martin [1971]). Such events are thus single points in time without internal structure. However, on the one hand, the PC signals an accomplished event (from the aspectual point of view) and thus conveys as its meaning the possible consequence of the event. The PS on the other hand is considered as aspectually unaccomplished and is used in contexts where time progresses and events are temporally ordered, such as narratives.

The IMP is imperfective (as its name suggests), i.e. it indicates that the event is in its preparatory phrase and is thus incomplete. In terms of aspect, the IMP is unaccomplished and provides background information, for instance ongoing state of affairs, or situations that are repeated in time, with an internal structure. Conversely, in English, the Simple Past (SP) is described as having as its main meaning the reference to past tense, and as specific meanings the reference to present or future tenses identified under certain contextual conditions (Quirk et al. [1986]). Corblin and de Swart [2004] argue that the SP is aspectually 'transparent', meaning that it applies to all types of events and it preserves their aspectual class.

In order to capture these EN/FR translation divergencies we have tried two different approaches. Firstly, by annotating the discursive feature of *narrativity*, we can disambiguate EN SP verbs toward the tense that should be used in FR depending on (non-)narrative contexts. Secondly, for the disambiguation of all verb types, we have experimented with an approach that directly uses, as labels, the FR tenses to which an EN verb should be translated. Classification and SMT with both approaches have been successfully implemented (Chapter 6 and Section 7.8.2).

As was argued above for discourse connectives, we operated with features specific to translation for verb tense as well (Section 6.2.1), aiming at fully functional SMT system pipelines

3. Consider for example the French sentence: *L'Union ne peut pas avoir comme objectif principal de réduire le niveau global des aides.* and its English translation: *The main objective of the Union cannot be to reduce the overall level of aid.*, where *L'Union* is subject in French, but becomes an attached PP to the *objective* in English, which there functions as subject of the sentence (Samardzic et al. [2010]).

rather than adapting more detailed or theoretically more grounded annotation frameworks.

3.3 Statistical machine translation (SMT)

In the following, we provide details on how statistical MT systems have developed into the dominating models that provide ease-of-use, speed and accuracy advantages over the previous rule-based implementations (Section 1.1).

The first statistical translation models were the so-called IBM models 1-5 (Brown et al. [1993]). The translation probability is defined as $p(e|f)$, i.e. the likelihood of a foreign string f to be a translation of a source language string e. During a so-called training stage, the goal is to approach a local maximum of the likelihood of a particular set of translations that is called training data. Models 1-5 are of increasing complexity in how the source and the target language words are aligned. In Model 1 only direct alignments are possible and each string length is considered to be equally likely. In Models 2-5 word reordering, link and word chunk dependencies are factored in, which is why more likely alignments are favored. Nevertheless, in all five models, the basic units that are aligned are words – which results in insufficient translation quality.

Phrase-based Statistical Machine Translation (PBSMT) (Koehn et al. [2003]) was the first significant improvement over the word-based statistical translation models and is, with some modifications to their decoding methods (Section 3.3.1), still at state of the art performance. The basic idea is to directly translate multi-word units; each source phrase is translated to a target phrase, with possible reorderings involved. For this, a translation table that maps not only words but phrases is built from the training data (aligned pairs of sentences). The term 'phrase' hereby does not refer to necessarily meaningful or linguistically motivated multi-word sequences, but to any chunk of words that can be seen in an entire sentence. A phrase like *fun with the* might be useful, e.g. for finding the correct German translation *Spass am* where the *am* is a contraction of the preposition and article (*an dem*) ([Koehn, 2010, p. 128]).

The main benefits of PBSMT models are the following ones. First, with phrases as atomic units, there are more one-to-one mappings, as opposed to word units, where there are frequent one-to-many mappings that may not be learned well. Second, word groups help to resolve translation ambiguities by modeling local dependencies (such as agreements between a noun and an adjective). Third, the phrase-based model is conceptually simpler than word-based models and makes more sense than the latter: arbitrary adding and dropping of words is not allowed.

3.3.1 Mathematical definition of phrase-based statistical machine translation

The formal framework of PBSMT starts by defining the best translation e_{best} as an argmax for the foreign input sentence f: $e_{best} = \arg\max_e p(e|f)$. The Bayes' rule is then applied

to separate the translation and language models (Noisy-Channel Model), which causes the translation direction to be mathematically inverted to $\phi(\bar{f}_i|\bar{e}_i)$ (Koehn et al. [2003]).

The argmax formula can be decomposed into three components that contribute to determine the best phrase translation pair. The translation probability ($\phi(\bar{f}_i|\bar{e}_i)$) accounts for the foreign phrases to match the English words. The reordering model ($d(start_{(i)} - end_{(i-1)} - 1)$, d for 'distortion') accounts for the fact that phrases may be ordered differently in SL and TL. The target language model $p_{LM}(e)$ is of arbitrary n-gram arity and weighs the fluency of the translation output. The complete PBSMT model is thereby given as:

$$e_{best} = \arg\max_{e} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_{(i)} - end_{(i-1)} - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1})$$

Within this framework and based on a parallel and sentence-aligned corpus, the goal in training the SMT system is then to extract all possible phrase pairs (source and target phrases) that are consistent with the word alignment (but not all of which will necessarily be correct). The probabilities of such phrases have to be estimated. It is then counted how often a particular phrase pair is extracted from sentence pairs (a value noted as $count(\bar{e}|\bar{f})$). The translation probability $\phi(\bar{e}|\bar{f})$ is eventually measured by the relative frequency of the pair.

After the training step, the parameters of the translation model can be tuned, usually by so-called Minimum Error Rate Training (MERT) (Och [2003]). The MERT algorithm optimizes linear weights relative to n-best lists of possible translations generated from a separate development (or tuning) corpus. The latter is much smaller than the training corpus and only consists of a few thousand sentences.The randomized optimization iterates between optimizing weights and re-decoding with those weights to enhance the approximation to the best translation (Cherry and Foster [2012]). Optimization is usually based on a loss function and for SMT, this is most often the BLEU evaluation metric (or rather $1 - BLEU$) (see Section 4.3 and Chiang [2012]).

At testing time, the so-called 'decoding' is the construction of the output sentence as a sequence from left to right by incrementally computing the sentence translation probability with the mentioned feature scores in the phrase table and formula shown above. For decoding, a beam search including stacking, hypothesis expansion and pruning is run over the phrase translation table in order to guarantee computability and performance. A proper trade-off between speed (small beam size) and performance (large beam size) has to be found.

3.3.2 SMT models for using linguistic information

In the following, we present a series of methods and models with which linguistic information can be integrated into SMT. These methods comprise to better translate syntax, semantics,

word senses and discourse phenomena (pronominal anaphora, lexical cohesion, discourse relations) (this section), and verb tense (Section 3.3.3).

Phrase-based vs. hierarchical statistical machine translation

Along with phrase-based statistical machine translation, there are methods to integrate linguistic information into SMT. In addition to the linearly operating PBSMT, especially the *Hierarchical* Phrase Models are noteworthy. They combine the idea of the phrase-based models and of tree structures by using chart parsing for decoding (Chiang [2005]). Other models include explicit syntactic annotation, i.e. syntax trees, where the modeling can take place on the source language and/or target language side: tree-to-string models (Zhou et al. [2008]), string-to-tree (Zollmann et al. [2006]) or tree-to-tree translation (Nesson et al. [2006]). Hierarchical models do not perform necessarily better than phrase-based ones, as syntax trees and grammars might diverge too drastically in SL and TL. For this thesis we have therefore focused on PBSMT.

Factored translation models

Factored translation models (Koehn and Hoang [2007]) have been proposed as a general way to use additional knowledge within the SMT paradigm, possibly coming from text-level features. Factored models, as currently implemented in SMT toolkits such as Moses and cdec (Koehn et al. [2007], Dyer et al. [2010]), are most often used to add morphological information (e.g. to translate to a morphologically-rich language), but also semantic information.

Factored translation models with semantic information have been studied by e.g. Baker et al. [2012] who augment hierarchical, syntax-based translation models by adjoining semantic labels. The labels produced by named entity recognition, modality and negation taggers were appended to the nodes in the syntactic tree input, in order to build the translation models. As a result, Urdu/English translation was improved by 0.5 BLEU points over a syntax-only baseline.

Birch et al. [2007] made use of supertags in a Combinatorial Categorial Grammar as factors for translation models. When the supertags (combined with other factors, e.g. POS tags) were applied on the target language side only, the factored models improved over a phrase-based only model by 0.46 BLEU points for Dutch/English translation. However, when the factors were only applied to the source side, the factored models did not conclusively improve German/English translation. Wang et al. [2012] have shown improvements for BLEU and manual evaluation for Bulgarian/English translation when using as factors POS, lemmas, dependency parsing, and minimal recursion semantics supertags.

Due to simplicity of use and known capacity to deal with linguistic features in SMT, we will make use of factored translation models in the present thesis (see Section 7.7). Nevertheless, we will also present results of several other approaches, including new ones, that take

advantage of the discourse labels output by our classifiers (see Chapter 5).

Word sense disambiguation for MT

The disambiguation of discourse connectives can be seen as an instance of the word sense disambiguation problem. The two tasks are similar as one tries to find the sense which is signaled by a word in a specific context. In word sense disambiguation settings however, content words only are considered and these can sufficiently be disambiguated with n-gram features. Section 5.2 will show that for discourse connectives (a class of function words), more elaborate and longer distance features are needed to reach the same disambiguation performance.

The word sense disambiguation methods however are useful and related to the methods described in this thesis, as they have been applied to both, function words and integration into SMT, which is the topic of the following subsections.

With word sense disambiguation methods for content words, Chan et al. [2007] as well as Carpuat and Wu [2007] obtained slight translation improvements. The latter authors used the translation candidates output by a baseline SMT system as word sense labels. Then, the output of several classifiers based on linguistic features was weighed against the translation candidates output by the baseline SMT system. Therefore, integration of MT and WSD amounted to *postprocessing of MT*, while in the present proposal, connective labeling amounts to *preprocessing*. The WSD+SMT system of Carpuat and Wu [2007] improved BLEU scores by 0.4–0.5 for English/Chinese translation. Xiao et al. [2011] identified ambiguous words in the SMT system output and then re-decoded using a filtered set of translation options (e.g. using the most frequent translation), focusing on document-level consistency.

Word sense disambiguation methods for function words (the word class connectives belong to as well) have been rarer and translation models integrating these are even less studied. Chang et al. [2009] disambiguated the Chinese particle 'DE' which has five different context-dependent usages (modifier, preposition, relative clause, etc.). When the linguistically-informed LogLinear classifier was used to label the particle prior to SMT, the translation quality was improved by up to 1.49 BLEU points for phrase-based Chinese/English translation. Similarly, Ma et al. [2011] proposed a Maximum Entropy model to annotate English collocational particles (e.g. come *down/by*, turn *against*, inform *of*) with more specific labels than a standard POS tagger would output. Such a tagger could, as the authors suggest, be useful in the future for English/Chinese translation.

A number of papers have studied the hypothesis of 'one sense per discourse' in the case of MT (Carpuat [2009], Carpuat and Simard [2012]), finding that using only one translation per discourse (i.e. translating a source word via the same target word in all of its occurrences in a text) can improve BLEU scores when using supervised WSD. For instance, Xiao et al. [2011] identified ambiguous words in the SMT system output and then re-decode using a filtered set

of translation options (i.e. using the most frequent translation), focusing on document-level consistency (but their method is difficult to extend to other discourse-level phenomena).

Integrating WSD with MT raises decoding problems (due to the larger search space) which do not apply to discourse connectives. In fact, most WSD methods either rely on very local criteria that could be learned by current phrase-based SMT models, without the need for additional processing, or on global text-level topics – for which attempts to integrate them with MT already exist (Eidelman et al. [2012]).

Text-level and discourse information in SMT

The significance of discourse information has long been acknowledged for MT, but using such information remains a major challenge for implementation into operational systems, be they statistical or rule-based.

As early as 1999 (Mitkov [1999]), there were several proposals on how to integrate the resolution of referential anaphora into MT. Anaphora such as referential pronouns remain a big challenge for MT as current models most often are still limited to sentence-based translation, which is why knowledge about gender and number of an antecedent will be lost for a pronoun or referential expression in the current sentence. With anaphora resolution being itself a difficult NLP task, the proposals were as broad as using rule-based resolution, document- or topic-constraints or full syntactic parsing in order to resolve anaphora prior to MT (Mitkov [1999]).

For SMT, several methods have been proposed during the last years to constrain pronoun choice (Hardmeier and Federico [2010], Le Nagard and Koehn [2010], Guillou [2012]), relying on knowledge of their antecedent, which is imperfect due to anaphora resolution errors. In a more syntactically oriented approach, Novak et al. [2013] built an English/Czech translation system that relies on rich syntactic annotation, external anaphora resolution tools and lexical co-occurrence features in order to better translate the English genderless pronoun *it* into Czech.

For the translation of entire discourse structures at the paragraph level, an early proposal by Marcu et al. [2000] anticipated the architecture of a discourse-aware MT system for English/-Japanese, which are languages that organize discourse very differently. Such a system would consist of the following three modules:

1. a discourse parser, that e.g. derives, for both languages, the discourse tree in RST-like manner as described above
2. a discourse structure re-writing module that renders the Japanese discourse tree closer to the English one, by re-ordering rules
3. an SMT system including a language model that would incorporate features from the discourse structure trees

These three steps are inspired from and also necessary when integrating syntactic information into SMT models (as described above). Marcu et al. however only focused on a feasibility study for module 2 and left MT experiments with module 3 for future work, which has, to our best knowledge, not been implemented so far. We have not implemented translation via discourse trees neither, given that phrase-based SMT systems still outperform hierarchical or syntactical ones, as has been explained above.

Lexical chains have only recently been considered for MT, in preliminary studies (Ture et al. [2012], Voigt and Jurafsky [2012]), showing the importance of referential cohesion.

As an alternative to current phrase-based, syntax-based and/or factored translation models, a text-level decoder for SMT, named Docent, was presented by Hardmeier et al. [2012]. Docent considers translation as an optimization task and allows for document-wide features. It was shown to perform the same as PBSMT when using standard translation features and allows for additional document-wide translation feature functions. But stacking information from previous translations in a document raises very large search space and efficiency issues, which was a further reason for trying to integrate our discourse-level features into standard PBSMT models.

A journal article summarizes most of the work on SMT with the broader perspective of discourse, lexical cohesion and co-reference in recent years (Hardmeier [2013]).

3.3.3 Verb tense in SMT

Features for verb tense, aspect and temporal connectives have been considered for natural language generation and interlingua-based MT in (Dorr [1992]) and (Dorr and Gaasterland [1995]).

Modeling verb tenses for SMT has only recently been addressed. For Chinese/English translation, Gong et al. [2012] built an n-gram-like sequence model that passes information from previously translated main verbs onto the next verb so that its tense can be more correctly rendered. Tense is not marked morphologically on verb forms in Chinese (where neighboring particles indicate tense), unlike in English, where the verbs forms themvselves are modified according to tense (among other factors). With such a model, the authors improved translation by up to 0.8 BLEU points.

Conversely, in view of English/Chinese translation but without implementing an actual translation system, Ye et al. [2007] used a classifier to generate and insert appropriate Chinese aspect markers that in certain contexts have to follow the Chinese verbs but are not present in the English source texts.

For translation from English to German, Gojun and Fraser [2012] reordered verbs in the English source to positions where they normally occur in German, which usually amounts to a long-distance movement towards the end of clauses. Reordering was implemented as rules on

syntax trees and improved the translation by up to 0.61 BLEU points.

For this thesis, similar to the handling of discourse connectives, we will make use of classifiers to automatically annotate the training data for SMT with labels that help to resolve the most urgent translation divergencies for tense and the EN/FR language pair.

4 Data, annotation procedures and evaluation metrics

In this chapter, we describe the two main data resources we used for the disambiguation of discourse relations and verb tenses as well as for training SMT systems: the Penn Discourse Treebank (Section 4.1.1) and the Europarl corpus (Section 4.1.2). Both needed some preprocessing to make them usable in our work. From Europarl, we selected translation pairs that included only source sentences that had been uttered by speakers of the source language and were then directly translated into the corresponding target language, i.e. without the translation being possibly distorted by a third-party or pivot language [1]. Preprocessing the PDTB facilitated feature extraction for automatic disambiguation, while processing the Europarl corpus provided us with translation pairs, where we could be sure that the discourse connective occurrences followed a 'natural distribution' and were not the effect of additional languages involved in the translation process. Besides Europarl, we used a few other parallel corpora in certain language settings and for tuning and testing (Section 4.1.3).

We will also present in this chapter the description of manual annotation experiments, both for connectives (Section 4.2.1) and verb tenses (Section 4.2.2), focusing on the definition of manual annotation procedures, the annotation granularity necessary for translation and automatic disambiguation, and the assessment of the obtained resources [2].

In all manual annotation experiments, we faced difficulties with discourse annotation that can be time-consuming and a challenging task for human annotators. In order to obtain

1. This procedure has been designed and subsequently published (Cartoni and Meyer [2012]) together with Bruno Cartoni, postdoc in COMTIS at the Linguistics department in Geneva.

2. For connectives, we largely could rely on and profit from collaboration in COMTIS with Sandrine Zufferey (researcher in linguistics at Geneva, at the time) and Bruno Cartoni (postdoc in linguistics at Geneva, at the time). The papers that are concerned, at least partly, with manual annotation of connectives and to which the author of the thesis contributed, were the following ones: (Meyer et al. [2011], Popescu-Belis et al. [2012], Cartoni et al. [2013b]). For verb tense and the annotation of narrativity, we closely collaborated with a COMTIS PhD student in Linguistics in Geneva, Cristina Grisot, on verb tense translation divergencies for the English/French pair (Meyer et al. [2013]) and (Grisot and Meyer [2014]). For the annotation of French verb tense translations onto English verbs, we co-supervised (together with Andrei Popescu-Belis) an intern at Idiap, Sharid Loáiciga, who implemented the semi-automatic annotation method and the oracle SMT experiment described in Section 4.2.2 (Loaiciga et al. [2014]).

reliable resources for the training of classifiers and MT systems, we defined new annotation methods and instructions, such as translation-spotting (Section 4.2.1) or semi-automatic methods (Section 4.2.2). After several rounds of annotation and after evaluation and clustering for the right granularity of the labeled connectives and verb tenses, we consolidated the resources and made them publicly available at www.idiap.ch/dataset/Disco-Annotation and www.idiap.ch/dataset/Tense-Annotation, respectively.

This chapter ends with a description of the evaluation metrics (Section 4.3) that we have used to measure performance of automatic classification and translation [3]. The metric relies on decisions whether an FR or DE connective is a valid equivalent to the EN one. The metric was evaluated in order to see whether its automatic scoring correlates with human judgments. The latter was indeed the case (with a small error range of about 2%) and we used the metric to score the translations output by our discourse-aware SMT systems described in Chapter 7.

4.1 Data

4.1.1 The Penn Discourse Treebank

The Penn Discourse Treebank (PDTB), version 2 (Prasad et al. [2008]), constitutes the largest manual annotation effort for discourse structure to date. It provides a separate annotation layer over the Wall Street Journal corpus, and contains the same WSJ sections (00-24) as in the Penn Treebank, a resource for syntactic annotation with hand-labeled syntactical trees (Marcus et al. [1993]).

In contrast to other existing resources for discourse, the PDTB follows a theory-neutral approach, in the sense that only the text spans or so-called arguments which are linked by a discourse connective are annotated, and not entire discourse structures (sometimes represented as trees over paragraphs, as in RST, see Chapter 3). This not only has the advantage to facilitate the annotation task but also guarantees interoperability with other annotation efforts.

Indeed, the PDTB approach has been adopted to annotate resources in other languages – however, this was done on different texts from the English PDTB, so that no parallel or translated version exists, with the exception of the Czech PDiT, see Section 7.1.1). Here, we first list the resources for the languages studied in this thesis, and then those for other languages.

- English
  - Discourse Graphbank (Wolf and Gibson [2005])
  - RST Discourse Treebank (Carlson et al. [2002])
  - Biomedical Discourse Relation Bank (BioDRB) (Prasad et al. [2011])

3. Najeh Hajlaoui (a COMTIS postdoc at Idiap at the time), was responsible for defining and implementing the metric for the evaluation of discourse connective translation. The author of the thesis contributed to parts of this metric, mainly to the form and granularity of the French and German dictionaries for discourse connectives.

- French
  - Annodis Corpus (Péry-Woodley et al. [2009])
  - French Discourse Treebank (FDTB) (Danlos et al. [2012])
- German
  - Potsdam Commentary Corpus (Stede [2004])
- Czech
  - Prague Discourse Treebank (PDiT) (Poláková et al. [2013])

Apart from these, there also have been efforts in other languages: Arabic (Alsaif [2012]), Chinese (Zhou and Xue [2012]), Turkish (Zeyrek et al. [2010]), Hindi (Kolachina et al. [2012]).

The PDTB differentiates two principal types of discourse relations: explicit and implicit. The former are expressed by about 100 English discourse connectives that have a lexical surface form, such as *although, however, meanwhile, since*, etc. Implicit relations have been annotated for cases where the annotators could infer a discourse connective that could be placed between two arguments. Often temporal or causal discourse relations can be inferred easily by the mere ordering of events described in a text, but we will also see that more complex relations like CONCESSION can be implicit (Section 8). For the entire WSJ corpus of about 1,000,000 tokens there are 18,459 instances of annotated explicit connectives and 16,053 implicit relations in the PDTB.

Connectives (and implicit relations) have two propositional arguments: the second argument is the one containing the explicit connective (or the one inferred by the annotators), while the first one is the linked span. The arguments and their spans are annotated as well. Annotators were asked to choose only the minimal amount of length required from the context to infer the discourse relation expressed by the connective or the implicit relation. Along with the arguments, several other features are annotated, such as information on polarity and whether an argument is part of an utterance by a third-speaker party.

Discourse relations, in PDTB terminology, are often also called 'senses' (of the connectives) and we too will use these terms interchangeably in this thesis. The PDTB organizes its set of 43 senses in a hierarchical way: there are 4 top-level discourse relations, followed by 16 sub-senses on the second hierarchy level and a further 23 senses on the most detailed third level (see the full hierarchy in Figure 4.1).

A hierarchical relation or sense structure has advantages over flat label sets that sometimes are established by adhering to certain discourse theories. The PDTB approach allows for specifying the level of detail necessary for the task at hand, e.g. even at creation time of the corpus, annotators were allowed to only insert a relation from the top four classes when they could not conclude on a more detailed relation from the subsenses. In addition, disagreements can be resolved, as was done in the PDTB, by moving up one level in the hierarchy (see 4.2).

A hierarchy also guarantees interoperability of the annotation, i.e. when there are different levels of granularity of the discourse relations, a mapping from one set of relations to another

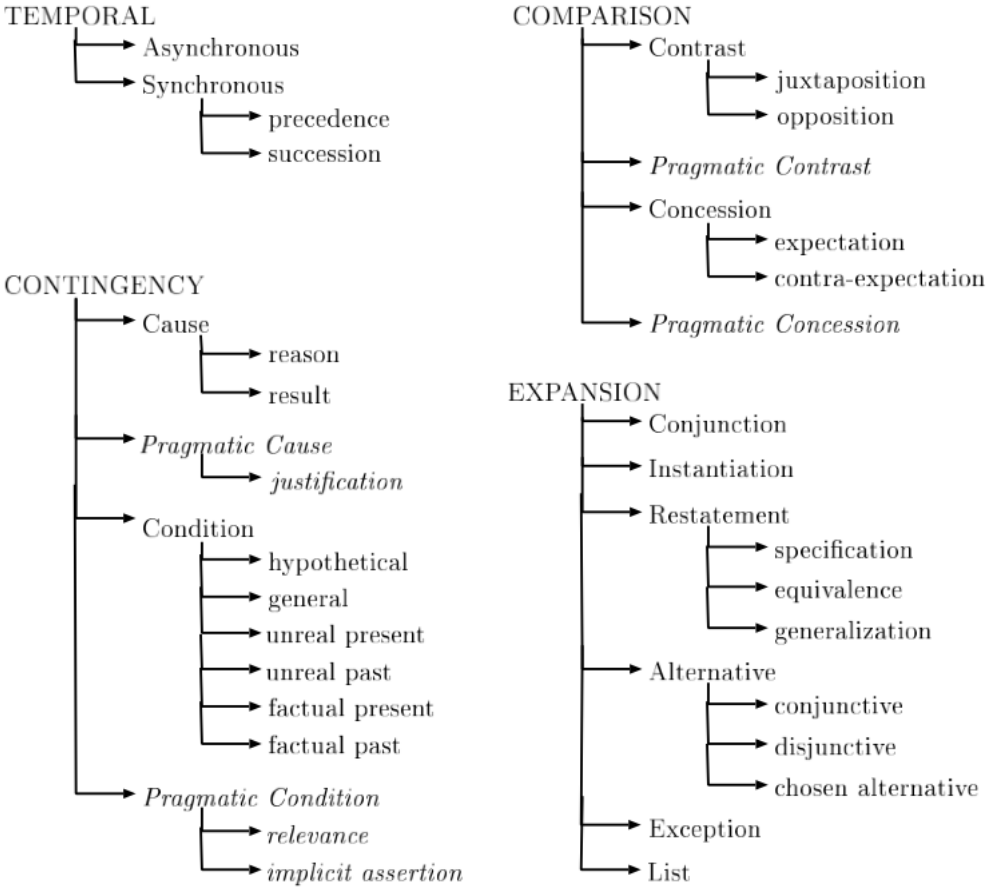


Figure 4.1: The hierarchical set of discourse relations (or senses) of the Penn Discourse Treebank, divided into three levels of detail: 4 top-classes with 16 sub-senses and 23 further sub-senses (taken from the PDTB annotation manual, Prasad et al. [2007], p. 27).

is more feasible, as it might be applied to only one sense level, or might even mix various levels of detail depending on the set of labels to be mapped to.

For feature and relation extraction we made use of an XMLization toolkit (Xuchen et al. [2010]) and a CSV version kindly provided by Christopher Potts [4].

4.1.2 The Europarl corpus

The Europarl corpus is one of the main resources used for machine translation and translation studies. The corpus consists of parallel texts of records of the debates of the European Parliament and was collected by the organizers of the annual Workshop on Machine Translation (WMT) who made it freely available (Koehn [2005]) [5].

4. http://compprag.christopherpotts.net/pdtb.html, we would like to thank these authors for making these tools available as they provide a lot of advantages over the PDTB native text file format.

5. www.statmt.org/europarl

For this thesis we have made use of versions 5, 6 and 7 of the corpus, the latter consisting of up to 23 languages according to the member states of the EU and debates from the years of 1990 to 2011. Tools are delivered with the corpus with which data from a source language can be aligned, sentence-by-sentence, to a target language, resulting in a parallel text with translation correspondences.

However, in the EU Parliament, each deputy has the allowance to speak in his mother tongue. As a result, when aligning a language pair, one has no guarantee that the language one considers as source actually is recorded in that language, whether it has been translated, or even been indirectly translated (via a pivot language).

Not paying attention to this fact can distort translational or contrastive linguistic studies, but can also actually decrease quality of the output of an MT system. Ozdowska [2009] has shown preliminary results where FR/EN SMT systems trained on direct FR/EN translation units performed marginally better than when pivot and third language FR/EN translations were mixed into to the parallel data. This was recently more thoroughly confirmed by Lembersky et al. [2013] who treat the translationese problem as one of domain adaptation, where in-domain data is stemming from direct translations of the language pair and out-of-domain data is consisting of all other translations in Europarl. Out-of-domain data can still be fruitfully integrated when interpolated with the in-domain translation models. We additionally found in earlier studies and our own analyses of discourse connectives in parallel corpora, that there is considerable variation regarding the occurrence and usage of these elements among different languages (see Chapters 2 and 8).

The Europarl corpus provides meta-information for each statement, such as the speaker name and the language spoken by the parliamentarian. Thanks to this information, one can actually extract from the corpus all statements that were originally uttered in French, in German, etc., and their translations (after sentence alignment). These language tags are however not consistently present in all files for each language. In the following we are looking at the years 1990–2003 portion of the corpus. For these years of debates we know from personal discussion with a translator at the European Parliament that pivot translation was not used at all, which ensures directionality of all language pairs in the corpus. This is the case at least for frequent language combinations (there are fewer translators translating from Danish to Portuguese than from English into French). After 2003, the pivot language of English has been used in the translation process which implies that all statements were first translated into English and then into the 22 other target languages.

Table 4.1 provides figures for the language tags and years 1990–2003. As shown, only 66.53% of the statements contain a language tag. When comparing the files in different languages, a language tag is sometimes inconsistent, i.e. it can be present in the text file of one language but not in the corresponding file for another language (we counted 6619 such divergencies). As a consequence, in total, only 118’289 statements have a proper language tag.

In order to limit the effect of missing and diverging language tags, we preprocessed the corpus

Number of statements (in all languages)	187'720
Number of LANGUAGE tags	124'908
Number of diverging tags	6619
Remaining trustworthy tags	118'289

Table 4.1: Language tags in the Europarl corpus, years 1990–2003

to add and disseminate the language information to all files before extracting what we call 'directional sub-corpora', i.e. corpora where it is known that the source language was actually spoken in that language and that it has directly been translated to the other target language to extract. The correction and dissemination of (missing) language tags allows to (i) identify (and sometimes correct) the diverging tags – see Table 4.1, and (ii) to increase the number of statements in each directional pair. Table 4.2 below shows the increase in terms of number of statements for the English → French directional corpus.

Number of statements *before* dissemination	19'903
Number of statements *after* dissemination	24'725
Improvement	24%

Table 4.2: Improvement after dissemination/correction of language tags in the Europarl corpus.

With this procedure one can obtain as many directional corpora as there are language pairs in the corpus and the resulting directional subcorpora can be deployed in various translational or cross-linguistic studies. More details are given in two publications (Cartoni et al. [2013a]) and (Cartoni and Meyer [2012]).

Variation and features specific to source and translate language in terms of connectives will be addressed in Chapter 8. In classification (Chapter 5) and translation experiments (Chapter 7) we made use of both the directional sub-corpora as well as the entire Europarl corpus and will mention this along with the corresponding experiments.

4.1.3 Other corpora used for statistical machine translation

For English/Arabic translation and in certain English/French test configurations, for comparison purposes, we make use of the United Nations corpus (Rafalovitch and Dale [2009], see also Chapter 7, Sections 7.3 and 7.7.2). This is a parallel corpus with six languages (Arabic, Chinese, English, French, Russian, and Spanish), containing high quality translations of the resolutions of the UN General Assembly, with a size of about 3 million tokens per language.

Besides this corpus, which shares with Europarl a similarity of political vocabulary, we also make use of a large collection of newswire texts as distributed by the annual Workshop on Machine Translation (WMT). These are collections of news articles in their original language

that have been translated by humans into the other languages of the collection: Czech, English, French, German, and Spanish, sometimes also Italian and Hungarian. These sets are generally used as tuning and test sets at the WMT workshops, also due to their genre differences from Europarl, which is generally used for training. In our experiments, we employed the WMT news collections from the years 2008 to 2012. More details about these data sets are available at http://matrix.statmt.org/test_sets/list.

For English/Arabic translation, the tuning and test sets were taken from the annual NIST OpenMT evaluations. Similarly to the above-mentioned sets, this data consists of human translated newswire articles (see https://catalog.ldc.upenn.edu/LDC2013T07).

4.2 Annotation procedures

Human annotation at the discourse level was shown to be a difficult endeavor that relies on thorough instruction and training in the discourse units to be annotated. Inferring the discourse relations, even when signaled by explicit connectives, needs a high cognitive effort and still can result in rather low inter-rater agreements.

Discourse phenomena can in many cases be intuitively understood and correctly processed by readers, but when naming the relations and argumentative structures, difficulties arise depending on the granularity of the label sets used. Annotators might disagree on items (or item boundaries) to annotate, when either specific linguistic or world knowledge must be used in order to find the correct label. In the following two sections (4.2.1 and 4.2.2), we introduce the discourse phenomena for which we have performed manual annotation in order to obtain gold-standard resources on which the automatic classifiers can be trained and evaluated against.

4.2.1 Discourse relations

The automatic disambiguation of discourse connectives is usually approached as a supervised classification problem, where machine learning classifiers are trained over manually labeled data sets, which offer a gold-standard annotation for training and testing.

In the PDTB corpus, the annotators were provided with a sense hierarchy from which they could choose among the 129 possible senses and combinations across different sense levels of the entire hierarchy. Eventually, when counting actual occurring senses and combinations, only 63 have been used by the PDTB annotators.

Although the sense hierarchy is very detailed, good inter-annotator agreement was reported, reaching up to 92% for the four top classes, which however drops to 'only' 77% for third level relations. It must however be noted that these numbers are those obtained after the resolution of disagreements by resorting to the next higher relation in the hierarchy in cases where annotators decided on a different sub-sense (Miltsakaki et al. [2008]).

Sense annotation vs. translation spotting

To experiment along the same lines as was done in the PDTB, we performed manual annotation of connectives with their senses in our work, too. As a first experiment, we performed, as in the PDTB corpus, direct sense annotation for connectives. Project colleagues, who already had a thorough understanding of discourse and pragmatic theories annotated sentences extracted from Europarl, each containing a potentially ambiguous discourse connective (*while* and *since* in EN, *alors que* in FR, see below).

The PDTB and other discourse-annotated resources are monolingual only. Our overall goal however is related to multilingualism and translation. For this reason, we performed manual annotation of connectives in a multilingual parallel resource: the Europarl corpus.

In addition, the PDTB hierarchy seemed too fine-grained given current capabilities for automatic labeling and the needs for translating connectives, we defined a simplified set of labels for the senses of connectives, by considering their usefulness and granularity with respect to translation, focusing on those that may lead to different connectives or syntactical constructs in the target language. The senses found are more detailed than the four top PDTB senses but not as detailed as the third level ones. On the other hand, our sense inventory is biased by the translation direction of EN/FR and would need to be reconsidered when translating to another target language.

There are two major ways to annotate explicit discourse connectives. The first approach is to label each occurrence of a connective with a label for its sense, similar to the PDTB hierarchy of senses. However, as shown among others by Zikánová et al. [2010], this is a difficult and time-consuming task even when the annotators are trained over a longer period of time. This is confirmed by the rather low $kappa$ scores resulting from the manual sense annotations as can be seen for each connective in a dedicated subsection below (paragraphs 4.2.1- 4.2.1).

The second approach to annotation, which is the one that we pursued further in this thesis, is based on so-called *translation spotting*. The theoretical idea behind translation spotting is that differences in the translation of an item can reveal semantic features of the corresponding source language item (Dyvik [1998], Noël [2003]). In these studies, translation is used to disambiguate some semantic features of content words in the source language. Behrens and Fabricius-Hansen [2003] convincingly showed that using translated data can help to identify the semantic space of the coherence relation of ELABORATION, conveyed with a single marker in German (*indem*) but translated in various ways in English (*when, as, by + ing, -ing*).

Of course, translated texts do not always faithfully reproduce the use of language in source texts as translation has a number of inherent features such as the increased/decreased use of cohesion markers, as was mentioned in Chapter 2, which is problematic for automatic translation spotting where the term has first been coined. Véronis and Langlais [2000] considered the automatic extraction of translation equivalents in a parallel corpus. In our experiments however, the translation spotting is done manually in order to get fully accurate reference data.

	English Sentence	French Sentence	Transpot
1	In this regard the technology feasibility review is necessary, **since** the emission control devices to meet the ambitious NOx limits are still under development.	À cet égard, il est nécessaire de mener une étude de faisabilité, **étant donné que** les dispositifs de contrôle des émissions permettant d'atteindre les limites ambitieuses fixées pour les NOx sont toujours en cours de développement.	étant donné que
2	Will we speak with one voice when we go to events in the future **since** we now have our single currency about to be born?	Parlerons-nous d'une seule voix lorsque nous en arriverons aux événements futurs, **puisqu**'à présent notre monnaie unique est sur le point de voir le jour ?	puisque
3	In East Timor an estimated one-third of the population has died **since** the Indonesian invasion of 1975.	Au Timor oriental, environ un tiers de la population est décédée **depuis** l'invasion indonésienne de 1975.	depuis
4	It is two years **since** charges were laid.	**Cela fait** deux ans **que** les plaintes ont été déposées.	P (cela fait X que)

Figure 4.2: Examples of parallel sentences with the English connective *since* and its translation spotting in French. In the fourth example, the translation is not an explicit connective, but a paraphrase.

Attempts to perform translation spotting automatically (Simard [2003]) have proven to be particularly unreliable when dealing with connectives: Danlos and Roze [2011] assessed the performance of TransSearch (Huet et al. [2009]), a bilingual English-French concordance tool that automatically retrieves the translation equivalents of a query term in target sentences, and found that for the French connectives *en effet* and *alors que*, the tool spots a valid English translation in only 62% and 27.5% of the cases respectively. Compared to the general performance of the TransSearch tool for the rest of the lexicon (around 70% of accurate transpots), these results are particularly low. Danlos and Roze [2011] suggest that one possible explanation is the important number of possible translations that can be found for connectives, ranging from no translation to paraphrases and longer syntactic constructions, which therefore are difficult to spot automatically.

In the first step in our translation spotting procedure, human annotators work on bilingual sentence pairs, and annotate the translation of each connective in the target language. The translations are either a target language connective (signaling in principle the same sense(s) as the source one), or a reformulation, or a construct with no connective at all. In a second step of the annotation, all translations of a connective are manually *clustered* by the experimenters to derive sense labels, by grouping together similar translations.

Figure 4.2 gives an example of an excerpt of parallel texts as we had distributed to the annotators, with the found translations filled in. Sentences 1 and 2 show examples of the English connective *since* where it has a causal meaning, which is directly evident from the French translations *puisque* and *étant donné que*. In the third example, *since* has a temporal meaning, expressed in French with the connective *depuis*. As mentioned in the introduction, sometimes there is no one-to-one correspondence of a source connective with a target lexical form, as in the above example 4 where *since* is translated as an entire French paraphrase, *cela fait*, with a temporal meaning as well. When such paraphrases are found in translation spotting, they

often cannot directly be clustered, as an annotator would have to go through these instances again to determine the exact sense signaled. We therefore discarded paraphrases from the gold-standard resources in order to reliably train our classifiers on instances where an explicit and lexical target connective has been found. We will however address the problem of paraphrase translations, specifically in Chapter 8.

Sense	%	**French connectives**
Concession	25.45	si (54), même si (33), bien que (26), s'il est vrai que (14)
Contrast	7.89	tandis que (39)
Contrast-Temporal	18.24	alors que (91)
Condition-Temporal	2.00	tant que (10)
Comparison-Temporal	1.40	pendant que (7)
Simultaneity	0.80	lorsque (4)

Table 4.3: Sense clustering and sense distribution as percentage for the English connective *while* after translation spotting.

The second step, done by the experimenters, consists of grouping the found translations (in several hundreds of such bitexts) to so-called sense clusters, which are illustrated in Table 4.3 for the English connective *while*. The latter is translated into the French connectives shown in the third column and expressing the six senses shown in the first column (over about 300 instances). Finally, a connective substitution test has to be performed, which can either be done by the experimenters or annotators in order to make sure that the grouped connectives are interchangeable in most of the contexts. This was done by questionnaires where one of the annotators went through sentences where we deleted the connective beforehand and where he/she had to fill the connectives suitable for the given context.

This procedure ensures that the found sense clusters are valid for the language pair on which they were determined, but also has the disadvantage that it has to be repeated when a new target language is considered. We however found that translation spotting provides a fast and reliable way to perform discourse connective annotation in new texts that is especially suitable for the MT task as the sense clusters are exactly at the granularity level needed in order to disambiguate the most problematic discourse connectives for SMT. Further details about the method are provided in the publications (Cartoni et al. [2013b]) and (Popescu-Belis et al. [2012]).

In the following, we exemplify our experiments with sense annotation and translation spotting for three discourse connectives, before summarizing the full sets of annotations produced. We identified the two English connectives *while* and *since*, along with the French connective *alors que*, as being particularly problematic for translation because they are highly multi-functional, i.e. they can signal several senses and sometimes even two senses at the same time. For *alors que*, LexConn, a French database of connectives (Roze et al. [2010]), contains examples of sentences where *alors que* expresses either a BACKGROUND or a CONTRAST relation.

For the English connective *since*, Miltsakaki et al. [2005] identified three possible meanings: TEMPORAL, CAUSAL, and simultaneously TEMPORAL/CAUSAL. For WHILE, even more senses are observed: COMPARISON, CONTRAST, CONCESSION, and OPPOSITION. In fact, in the PDTB, the connective *while* is annotated with more than twenty different senses or combinations thereof.

Annotation of *alors que*. This first manual annotation involved two experienced annotators who annotated *alors que* in 423 sentences that were originally authored in French. The two main senses identified for *alors que* are BACKGROUND (labeled B) and CONTRAST (labeled C), as in the LexConn database. Annotators were also allowed to use a label J if they did not know which sense label to assign, and a label D for discarded sentences – due to a non-connective use of the two words which were not filtered out automatically (e.g. *alors, que fera-t-on?*). The annotators found 20 sentences labeled with D, which were removed from the data. 15 sentences were labeled with J by one annotator (but none by both), and it was decided to assign to them the label (either B or C) provided by the other annotator.

The inter-annotator agreement on the B vs. C labels was quite low, showing the difficulty of the task: $kappa$ reached $\kappa = 0.43$, quite below the 0.7 mark often considered as indicating reliability (Cohen [1960] and Section 4.3).

There are two principled solutions to deal with the difficulty when occurrences were annotated with B by one annotator and with C by the other annotator. Firstly, a double-sense label B/C for sentences labeled differently by annotators (B vs. C) can be defined. Such a label reflects the difficulty of manual annotation and preserves the ambiguity which is genuinely present in these occurrences.

Secondly, for comparison purposes, a second solution is to annotate the connective via translation spotting as explained above. *Alors que* appeared to be mainly translated by the following English equivalents and constructs: *although, whereas, while, whilst, when, at a time when*. Through this operation, inter-annotator disagreement can sometimes be solved: when the translation clearly is a contrastive English connective (*whereas* or *although*), then the C label was assigned instead of B/C. Conversely, when the English translation was still ambiguous (*while, whilst*, or *when*), the experimenters made a decision in favor of either B or C by re-examining source and target sentences.

Annotation of *since*. For *since*, 30 sentences were annotated by four experimenters in a preliminary round, with a $kappa$ score of $\kappa = 0.77$, indicating good agreement, and, for *since*, the feasibility of sense annotation without resorting to translation spotting. Then, each half of 558 sentences containing *since* was annotated by different annotators with three possible sense labels: T for TEMPORAL, C for CAUSAL and T/C for a simultaneously TEMPORAL/CAUSAL meaning. Two datasets can again be derived from this manual annotation: the double sense label T/C can either be kept (to study the effects of a supplementary label) or be converted to

label C.

Annotation of *while*. The English connective *while* is highly ambiguous. In the PDTB, occurrences of *while* are annotated with 21 possible senses, ranging from CONJUNCTION to CONTRAST, CONCESSION, or SYNCHRONY. We performed a pilot annotation of 30 sentences containing *while* with five different experimenters and the sense labels COMPARISON, CONCESSION, CONTRAST and TEMPORAL, resulting in quite a low inter-annotator agreement of $\kappa = 0.56$.

We therefore decided to perform a translation spotting task only, with two experienced annotators fluent in English and French. The observed translations into French confirm the ambiguity of *while*, as they include several connectives and constructs, quite evenly distributed in terms of frequency: *alors que*, gerundive and other reformulations, *si*, *tandis que*, *même si*, *bien que*, etc.

The translations were manually clustered to derive senses for *while*, in an empirical manner. For example, *alors que* signals CONTRAST-TEMPORAL, which is also true for *tandis que*, although the latter tends to be CONTRAST only more often. Similarly, *même si* and *bien que* are clustered under the label CONCESSION, and so forth.

The results of translation spotting (see Table 4.3) show that at least CONTRAST, CONCESSION, and several temporal senses are necessary to account for a correct translation. These distinctions are comparable to the semantic granularity of the second PDTB hierarchy level. Details on the annotation for these three connectives have been presented in Section 4 of (Meyer et al. [2011]).

The same procedure of translation spotting exemplified above on three connectives, has been used for 7 other English and 3 French discourse connectives. First, their translations were determined, and then they were clustered into sense labels, providing us with gold-standard resources which have been made available for further research, and are presented hereafter.

Published gold-standard resources

Several types of English and French discourse connectives, among which the most ambiguous ones, have been processed, aiming at 200 occurrences or more per type, and results are shown in Table 4.4. These types were selected because they were described in monolingual studies as having multiple possible senses – e.g. in various dictionaries, the PDTB, or LexConn. When annotating them by translation spotting, the *a posteriori* senses were sometimes different from the principal *a priori* ones listed in the literature, and both lists are represented in Table 4.4. Some sentences were discarded due to non-connective uses or other problems due to the automatic extraction of the occurrences. A total of 3231 connectives (2514 English and 817 French), of 12 types (8 English and 4 French), have been annotated, as summarized in Table 4.4. The resources were presented in (Popescu-Belis et al. [2012]) and the data sets for

each connective are available at https://www.idiap.ch/dataset/Disco-Annotation [6].

Lexical items	A priori senses	A posteriori senses	N.S.	F.S.
EN CONNECTIVE			*Total EN: 2,793*	
as	preposition; connective: causal, comparison, temporal	preposition; connective: causal, concession, comparison, temporal	600	599
although	contrast, concession	contrast, concession	197	183
even though	contrast, concession	contrast, concession	212	190
however	contrast, concession	contrast, concession	418	418
meanwhile	contrast, temporal	contrast, temporal	131	130
since	temporal, causal	temporal, temporal-causal, causal	558	421
though	contrast, concession	contrast, concession	200	155
while	contrast, concession, comparison, temporal	contrast, concession, temporal-contrast, temporal-durative, temporal-punctual, temporal-causal	499	294
yet	adverb; connective: contrast, concession	adverb; connective: contrast, concession	509	403
FR CONNECTIVE			*Total FR: 817*	
alors que	contrast, temporal	contrast, temporal, temporal-contrast	423	366
bien que	concession	contrast, concession	55	51
dans la mesure où	condition, explanation	condition, explanation	175	150
pourtant	contrast, concession	contrast, concession	312	250

Table 4.4: List of created resources in English and French. N.S. stands for number of automatically-extracted sentences submitted to annotators, and F.S. for the number of final sentences retained. The *a priori* senses are based on the PDTB (for English) or LexConn (for French) labels, while the *a posteriori* ones, as explained in the text, were defined by clustering after translation spotting and are specific to this work. Two sense labels clustered with '-' reflect genuine sense ambiguities. For *as* and *yet* we also included their POS tags (preposition and adverb) as additional categories, because they frequently appear with a non-connective usage.

4.2.2 Annotation of verb tense

For the annotation of temporal information, TimeML [7] is a rich framework that has been used to provide, similarly to the PDTB, reference corpora with gold-standard annotations for

6. Besides French and English, we also performed translation spotting (but no further sense clustering and consolidation) in 400 sentences for each of the 5 German connectives *aber, jedoch, während, wenn, wie*, available upon request.

7. Markup Language for Temporal and Event Expressions, see http://timeml.org/site/index.html.

temporal expressions and markers, in the so-called English (Pustejovsky et al. [2003]) and French TimeBank [8], for example. These two corpora, however, do not contain the same news articles and are not parallel or directly usable for the MT task. Moreover, the complexity of the TimeML annotation language makes annotation expensive and not fully reliable, either by humans or by automated methods (Verhagen and Pustejovsky [2008]). We have made use of TimeML as a feature for the discourse connective and verb tense classifiers (see Sections 5.3 and 6.2.1), but not for manual annotation. Rather, we looked at specific properties of verb tense that help to resolve translation divergencies in parallel corpora, for the English/French language pair.

A first approach is to look at a prominent translation problem which consists of the English Simple Past (SP) tense that can be translated in French by at least three verb tenses (Passé Composé, Passé Simple, Imparfait). A binary discursive feature that helps to find these different usages of the EN SP is *narrativity* as it was defined above in Section 2.2, its manual annotation is described below.

A further possibility is to semi-automatically align English and French verb phrases, and to record the French verb tense each English verb was translated to. The predicted FR tense can be used as label onto the EN verb form.

Narrativity

A manual annotation experiment was conducted to empirically test if the narrative and non-narrative usages of the SP can reliably be detected in EN. Two EN native speakers went through a training phase in order to check whether the instructions given were clear. The annotators had to annotate 10 text excerpts where the SP occurred and to explain orally their reasoning.

The annotation guidelines included: (a) a definition of narrativity, (b) the explanation of each usage (narrative and non-narrative) with examples, (c) the instruction to read each excerpt, identify the verb highlighted and decide in context, the role of the highlighted verb and whether the connective *and then* could be added without changing the meaning (the verb would have a narrative usage) or not (non-narrative usage).

The data used for the annotation experiment was taken from the parallel corpus by Grisot and Cartoni [2012]. The sentences come from parallel EN/FR corpora of four different genres: literature, news, parliamentary debates and legislation. From this corpus, a subset of 458 excerpts (which we call items) containing occurrences of the SP was given to the two human annotators. For each item, the sentence with the SP verb, as well as one sentence before and after them, have been provided for sufficient contextual information.

The results of the human annotation experiments have been analyzed in three steps. As a first step, it can be tested whether different raters produced consistently similar results, so

8. https://gforge.inria.fr/projects/fr-timebank/

that one can infer that the annotators have understood the guidelines. In our annotation experiment, the two annotators agreed on 325 items (71%) and disagreed on 133 items (29%). This results in a $kappa$ value of 0.42, which is above chance, but not high enough to consider the annotation as reliable ($\kappa > 0.6$ or 0.7).

Error analysis revealed that the main source of errors was the length of the temporal interval between two eventualities perceived differently by the two annotators, which led to ambiguity between temporal sequence or simultaneity, corresponding to narrative and respectively non-narrative usage. This has been corrected in a second annotation round, where the insertion of a temporal connective was expected to force a narrative or a non-narrative reading. Disagreements were thus resolved in the second annotation round, with two new annotators, on a clean corpus containing 439 items. Annotators have been asked to insert a discourse connective in order to explicitate the implicit relation existing between eventualities. The connectives *and then/before* signaling temporal sequencing and *because/thus* for causal relations were proposed by annotators for the *narrative* label. For the *non-narrative* label, the connective *and* expressing simultaneity or no connective possibly inserted have been proposed. The inter-annotator agreement was 0.91, signaling very strong and reliable agreement. Here, only 4 items of disagreement were found, which were discarded from the corpus, which contains 435 items.

The data consisting of the items where the annotators agreed from both rounds has also been used for mappings of the EN SP against the tenses used in the target language FR, taken from the parallel corpus. The narrative usages identified by annotators correspond to translations by the FR tenses Passé Simple/Passé Composé and the non-narrative usages correspond to translations by Imparfait in 80% of the cases. This shows that narrativity is a reliable indicator of French past tense usage and only leaves 20% of cases where annotators agreed on the narrativity label but where there is no correlation with the tense used in FR (these instances can however still remain in the corpus, as there was actual inter-annotator agreement on the narrativity labels).

These manual annotation experiments have been presented in (Meyer et al. [2013] and Grisot and Meyer [2014]) and have illustrated that it can be difficult, even for humans, to infer from the context all the semantic features (such as narrativity), which in turn has effects on translation quality. The following approach uses semi-automatic methods in order to indicate directly for each EN verb phrase the FR tense it was translated to by a human translator.

Annotation of translated FR tense onto EN verbs

The automatic annotation of the FR tense used by a professional translator when translating an EN verb phrase has some advantages over using the binary narrativity feature presented above:

— the EN tenses need not be restricted to Simple Past

— the verbs and tenses to annotate can be extracted automatically
— no manual annotation and training is needed

Using state of the art tools (word alignment, dependency parsing, morphological analysis) and the Europarl parallel corpus, the FR verb tense an EN verb should be translated to (according to the human reference translation) can be found and aligned automatically.

Starting from the entire sentence-aligned Europarl corpus v7 for English/French (2'008'710 sentences), we make use of Giza++ (Och and Ney [2003]) to align the EN source text with the FR target at the word level. Additionally, we parse the EN side with a dependency parser (Henderson et al. [2008]) that outputs, for verbs, their categories such as VB (verb base), MD (modals) and VC (verb chain). For FR we make use of MORFETTE (Chrupała et al. [2008]), an automated morphological analyzer that makes hypotheses on the tense of verbs in a sentence, such as *V-indicatifpresent1p* for a verb in indicative present tense, first person, plural.

In a second processing stage, we use a set of hand-written rules to infer VPs and tense labels on the basis of the morpho-syntactic annotation, independently for both sides of the parallel corpus. For example, if two words in EN tagged as MD (Modal) and VB (Verb Base-form) are found, several tests follow: first, it is checked if MD is the head of VB, then if they are bound by the VC (Verb Chain) dependency relation. If this is the case, then the whole sequence (MD VB) is interpreted as a valid VP. Last, in this particular case, the first word is further tested in order to disambiguate between a future tensed verb or a "conditional construction". Conditional constructions comprise all VPs including a modal verb, apart from *will* and *shall* (i.e. *should, would, ought, can, could, may, might*). The voice (active or passive) is considered for both languages, because it helps to distinguish between tenses with a similar syntactical configuration (e.g., *Jean est parti* vs. *Jean est menacé*, meaning 'Jean has left' vs. 'Jean is threatened'). Indeed, while all forms of passive voice in French use the auxiliary ÊTRE (EN: *to be*), only a small set of intransitive verbs (recognized by our rules) use it in their compound forms. This example also illustrates the main reason for using MORFETTE for French parsing: it produces both morphological tagging and lemmatization, which are essential for determining the French tense.

We have observed 24 principal voice/tense combinations in EN and 24 in FR (i.e. 12 active forms and 12 passive forms for each). As a consequence, a core set of 24 rules was defined for each language, one for each tense in each voice. However, some verbs need further disambiguation. English conditional and future tenses need one additional rule to distinguish between them. Besides, French active compound tenses with the auxiliary ÊTRE are syntactically ambiguous, and two more rules were defined for disambiguation. This sums up to 25 rules for EN and 26 for FR. These rules are robust and for cases where the EN and FR parses are correct, the tenses can be inferred at full accuracy. Also depending on the parses, only VP pairs which are assigned a valid tense on *both* EN and FR sides are retained in the data set.

EN pos	EN words	EN tense	EN voice	EN POS	DI	EN dep	FR words	FR tense	FR voice
1	The	–	–	DT	2	NMOD	Des	–	–
2	same	–	–	JJ	3	SBJ	similaires	–	–
3	was	sim_past	passive	VBD	0	ROOT	ont été	**passe_comp**	active
4	said	sim_past	passive	VBN	3	VC	déclarations faites	**passe_comp**	active
5	of	–	–	IN	4	ADV	@	–	–
6	GATT	–	–	NN	5	PMOD	accord du GATT ceux escomptés	–	–
7	,	–	–	,	6	P	,	–	–
8	and	–	–	CC	6	COORD	mais	–	–
9	look	other	n/a	VB	8	CONJ	@	no_tag	n/a
10	what	–	–	WP	11	SBJ	résultats	–	–
11	happened	sim_past	active	VBD	9	OBJ	étaient contraire	**imparfait**	active
12	there	–	–	RB	11	LOC	@	–	–
13	.	–	–	.	0	ROOT	.	–	–

Figure 4.3: An EN/FR translation (columns marked 'EN words' and 'FR words') that was word-aligned with Giza++, parsed for dependency in EN and analyzed morphologically by MORFETTE in FR. The verb tenses (in bold) are then inferred by a small set of hand-crafted rules. 'DI' stands for the dependency index in the EN parse.

Published gold-standard resources

The three outputs from the tools can be combined and formatted as illustrated in Figure 4.3. Due to errors from each tool, be it alignment, parsing or morphological tagging errors, the automatic annotation procedure has a rather low recall in terms of labeled verbs with respect to the entire corpus (62% for EN and 42% for FR verbs). In terms of precision however, the procedure provides a highly reliable and reusable resource with correctly identified and labeled verbs in 97% of all cases for EN and 80% for FR (as found through manual assessment on a subset of the data).

In the total parallel EN/FR text, there are 419'419 annotated sentences and 454'890 annotated verbs in EN. Due to the errors and when only keeping instances where the EN labeled verb phrase is aligned to a valid FR tense, the published gold-standard resource amounts to 203'140 sentences with an average of 3.3 verbs per sentence. We set aside from this corpus 7000 sentences: 4000 for tuning and 3000 for testing the classification and SMT systems. The detailed statistics per tense class occurring in the corpus are given in Table 4.5[9]. The annotation method and statistics on results have been published in Loaiciga et al. [2014].

4.3 Evaluation metrics

In this section, we provide an overview of the (semi-)automatic scoring tools and metrics we made use of, on the one hand, to evaluate the performance of the classifiers for connectives and verb tenses, and on the other hand to evaluate the quality of the translation output by baseline and augmented SMT systems.

9. The corpus is freely available at: https://www.idiap.ch/dataset/Tense-Annotation

Tense	Training set	Tuning set	Test set	Total
Imparfait	9'561	135	122	9'818
Impératif	249	5	4	258
Passé Composé	42'112	754	636	43'502
Passé Récent	197	4	3	204
Passé Simple	465	9	6	480
Plus-que-Parfait	2'075	22	17	2'114
Présent	169'520	3'531	2'618	175'669
Subjonctif	4'597	71	78	4'746
Total	228'776	4'531	3'484	236'791

Table 4.5: Sizes of the training, tuning and test sets for French tense prediction and SMT, with statistics per tense.

Metrics for classification. The accuracy of connective disambiguation is rated, as in previous work, using classic accuracy (percentage of correctly classified instances), precision (correctly classified instances among correctly identified ones) and recall scores (correctly classified instances over all instances). When averaging over all classes, one obtains the F1 score ($F1 = 2 * (Precision * Recall)/(Precision + Recall)$). The score that we use is the weighted average of F1 scores taking into account the size of each ground-truth class (micro-averaged F1), or, when applying uniform weights per class, its macro-averaged variant. The same scores are used for evaluating the performance on disambiguating verb tense, as it is addressed here by similar approaches, i.e. as a supervised classification problem.

Apart from the F1 score, we also report, mostly for manual annotation experiments, the so-called $kappa$ value, which is an indicator of the reliability of the produced annotation. $kappa$ is computed over the agreements and disagreements of two ore more annotators (or between a gold standard annotation and a classification system output) and takes into account that some items might have been annotated just by chance or randomly (Carletta [1996]). $kappa$ values are in a range of -1 (complete chance) to 1 (complete agreement), with 0 to 0.4 considered as low agreement, 0.4 to 0.6 as reasonable agreement and 0.7 to 0.9 as high agreement.

Automatic scoring for MT. Automatic scoring of translation quality is a difficult problem and has become a research task in its own right over the last years (King et al. [2003], Koehn [2010], Chapter 8). This is mainly due to the fact that there is no single, one-best translation and that human reference translations differ considerably when several human translators provide translations even for just a short sentence.

The metrics most often referred to in the literature all rely on the same scoring principle: the overlap of a system's output (or candidate translation) with one human reference translation, or, depending on availability, several different reference translations. This overlap can be measured by various approaches: the BLEU score (Bilingual Evaluation Understudy, Papineni

et al. [2002]) for example, counts overlap in terms of matching n-grams, and is the most frequently used metric. The more matches there are for (usually) 4-, 3-, 2- and 1-grams in a candidate translation vs. its reference, the higher the BLEU score. The values of the score range from 0 to 100, where 100 is reached for identical translations. State-of-the-art systems, depending on the language pair involved, tend to have values between 11 and 33 BLEU points. Although criticized frequently for its limitations, BLEU remains a fast, language-independent and freely-available metric for MT, which correlates rather well with human judgments of translation quality, especially when averaged over a large quantity of text. Other frequently used measures are METEOR (Metric for Evaluation of Translation with Explicit ORdering, Denkowski and Lavie [2011]) and TER (Translation Error Rate, Snover et al. [2006]). The former considers possible word re-ordering and synonyms (with values similar to BLEU) and the latter computes a string edit distance in terms of word insertion or deletion that would be needed to transform a candidate into a reference translation (the smaller this edit distance is, the better the translations). For our task, we most often compare a modified, discourse-aware SMT system against a baseline system; we observed that BLEU, METEOR and TER scores show the same behavior of improving or degrading. This is why we will only report BLEU scores in the remainder of the thesis. When not stated otherwise, BLEU is computed via the NIST MTeval script v. 11b [10].

The design of an SMT system includes a tuning stage where feature weights are optimized in order to find the best translations. For most of the systems described in this thesis we use Minimum Error Rate Training (MERT, see Chapter 3, Section 3.3 above and (Och [2003])).

MERT is implemented as a randomized, non-deterministic optimization process, so that each run leads to different feature weights and as a consequence, to different BLEU scores when translating unseen text. One way to improve confidence in the BLEU scores, especially when test sets are small, is to bootstrap BLEU scores (Zhang and Vogel [2010]): the test sets are re-sampled a thousand times and the average BLEU score is computed from individual sample scores. Another way is to run MERT several times (usually 3 to 5), average the scores, and perform a t-test to compute p-values for the significance of the score differences. When these values are below 0.05, they confirm that it is statistically likely, that such differences would be observed in other tuning runs. This procedure is implemented in the MultEval tool, version 0.5.1 (Clark et al. [2011]). The BLEU scores within this tool are computed by jBLEU V0.1.1, a reimplementation of NIST's MTeval script in version 13 without tokenization, see footnote 10.

New MT evaluation metrics. Given the small range of changes to the discourse units dealt with in this thesis, it is likely that classical MT scoring is not sensitive to them, and would not make visible enough the improvement in their translation in terms of global score. One way to circumvent this problem is using manual evaluation of translations and counting how many correct and incorrect changes were output for a specific discourse phenomenon by an augmented SMT system vs. a baseline one.

10. Available from www.itl.nist.gov/iad/mig/tools/

We evaluated the newly built SMT systems described later in this thesis in this way, i.e. by considering a representative amount of test translations (usually around several hundreds) and counting the number of time the translation of a connective (or verb tense) by our modified models was better or comparable or worse than a baseline translation or a human reference translation. This method can be time-consuming but provides a precise assessment of the system's improvement in terms of translation quality and coherence.

A recently developed metric for discourse connectives (semi-)automatically compares the translations of connectives between a reference and a candidate translation. ACT, for Accuracy of Connective Translation (Hajlaoui and Popescu-Belis [2013]) [11], attempts to identify the translation of each source connective in a reference and a candidate translation using word alignment and several heuristics. The two translations are compared according to the following possible cases: identical (case 1); 'synonymous' according to a predefined, sense-specific dictionary (case 2); or incompatible in terms of connective senses (case 3). Moreover, the candidate connective can be missing (or possibly not identified by the alignment procedure, case 4), or the reference connective can be missing (case 5), or both (case 6). For each source connective, ACT scores one point for cases 1 and 2 (C_1, C_2, by number of instances), and zero for all others. The total score (named ACT_a for automatic) is then normalized by the number of source connectives (N), and ranges from 0–100, where 100 means that every single connective output by a discourse-aware system is the same as (or equivalent to) the corresponding one in the reference translation. The following equations formalize this first variant of ACT, along with two others: In ACT_{a5+6}, either all cases 5 and 6 (C_{5+6}) are excluded from the count, given that it is not automatically decidable whether they contain actually correct translations or not. This variant therefore always amounts to a higher score than the other two. Finally, in ACT_m (manual), the Cases 5 and 6 are judged manually (noted C_{5+6_corr}) in order to find the most accurate score that considers actually correct translations as precisely as possible, which is time-consuming because of the human effort.

$$
\begin{aligned}
ACT_a &= (|C_1| + |C_2|)/N \\
ACT_{a5+6} &= (|C_1| + |C_2|)/(N - |C_{5+6}|) \\
ACT_m &= (|C_1| + |C_2| + |C_{5+6\_corr}|)/N
\end{aligned}
$$

ACT was shown to be within 2-5% of human scores on the four target languages used in the thesis (French, German, Italian, Arabic). The ACT metric can be ported to other linguistic phenomena such as verb tense and pronouns. In the experiments on verb tense translation however, we did not attempt to design and validate such a new metric, but rather resorted to manual evaluation along the lines described above. Therefore, we counted how many translations generated by a tense-aware SMT system would be better, equal or worse compared to a baseline, in terms of verb tense, lexical choice and overall correctness of the verb phrase translation.

11. ACT is available under GPL v3 license from: https://github.com/idiap/act.

5 Automatically disambiguating discourse connectives

This chapter is dedicated to automatic disambiguation methods for discourse connectives. The training of the machine learning classifiers relies essentially on the manually annotated datasets that have been described in the previous chapter. We will first introduce the algorithms that have been used to disambiguate connectives, in the state of the art and in our work (Section 5.1). This is followed by an experiment in which we have compared the disambiguation of connectives to word sense disambiguation. The two tasks are similar as one tries to find the sense which is signaled by a word in a specific context. In standard word sense disambiguation settings however, content words only are considered and these can sufficiently be disambiguated with n-gram features. Section 5.2 shows that for discourse connectives, more elaborate and longer distance features are needed to reach the same disambiguation performance. What these features are and how we extracted them from our data is then described in Section 5.3.

Initially, classifiers were trained on PDTB data, and results of these experiments are presented in Section 5.4 (see also Meyer [2011] and Meyer and Popescu-Belis [2012]). Although these classifiers make use of the state of the art features and perform at F1 scores between 0.5 and 0.9, applying them to Europarl data (for building SMT systems) is problematic because of the genre change from newswire text (in the PDTB) to parliamentary debates. This can lower performance of automatic classification. As soon as the first connectives were manually annotated we therefore also started disambiguation experiments with Europarl data, reporting results in Section 5.5 (see also Meyer et al. [2011], Meyer and Popescu-Belis [2012], and Meyer et al. [2012]). New features, especially semantically-oriented and translational ones, helped to increase performance for highly ambiguous connectives and to advance the state of the art for these connective types. Section 5.6 furthermore presents experiments where we combined PDTB and Europarl data (via sense label mapping) in order to have more training instances. Cross-validation experiments and feature analysis reveal that for all the 7 connectives to classify, the performance can reach the human agreement level for the second level of the PDTB hierarchy of senses (F1 scores of 0.7–1.0 depending on the connective and feature selection).

As a final point in Section 5.6 we show that the distribution of connectives in the test sets can affect the overall performance: not all connectives are equally difficult to classify and we will present an analysis which also points to translation performance (Chapter 7, Section 7.7.2), which is directly affected by connective classification performance.

5.1 Algorithms

As with many other classification tasks in NLP, the disambiguation of connectives usually is addressed as a supervised learning problem where algorithms make use (at training stage) of the information provided by hand-labeled data. This is due, on the one hand, to the specific context features that are needed to find discourse relations and on the other hand to the fact that the few studies on unsupervised disambiguation have all reported lower performance than supervised ones (Pitler et al. [2009], Lin et al. [2009], Zhou et al. [2010]).

In the following sections of this chapter, we will compare Random Forests, Naive Bayes, Support Vector Machine, Conditional Random Field and Maximum Entropy algorithms, most of which have been used in previous work for connective disambiguation. Here, we briefly illustrate the *a priori* advantages and drawbacks of these algorithms for the task under study and draw some initial arguments for using Maximum Entropy to label our data for SMT.

Decision Trees and Random Forests Decision Trees (such as, for example, the C4.5 algorithm proposed by Quinlan [1993]), or an ensemble of them, a so-called 'Random Forest' (Breiman [2001]), have the advantage that they can easily be visualized to see which features actually contribute most to solve the classification task. Most often however, the decisions are binary only, as they are based on a yes/no decision for a specific feature without considering all features at that decision point.

Naive Bayes Naive Bayes classifiers were among the first algorithms to be successfully used for the disambiguation of connectives (Pitler and Nenkova [2009]). A comparison with a maximum entropy algorithm in that work did not yield better performance.

Support Vector Machine SVMs have been used for a large range of machine learning problems and perform well because they can linearly (in the feature space) separate non-linearly separable data thanks to the use of kernels that project the data onto an implicit, higher dimensional space. SVMs are for example part of the discourse parser designed by duVerle and Prendinger [2009]: the authors mention that SVMs overcome generalization errors (overfitting) and can be trained over a large set of features. We also successfully applied SVMs (in the implementation of Chang and Lin [2011] and Hall et al. [2009]) for discourse connectives. In our configurations, the maximum entropy algorithm (see paragraph below) however outperformed the SVM-based classifiers.

Conditional Random Fields CRFs (Lafferty et al. [2001]) are suitable for sequence-labeling tasks such as POS tagging, where normally only a few preceding words and tags are needed to find the current one. This does not necessarily hold true for discourse connectives or discourse relations, although they can appear in a sequence – which is the reason why we also experimented with CRFs. Most often however, features from a wider context are needed to find a specific, possibly ambiguous relation. If such features from a wider context were integrated in a CRF, the model would become difficult to train, due to longer label sequences.

Maximum Entropy Maximum entropy models are discriminative and based on conditional probabilities that can be calculated from the class distribution present in the data. The name 'maximum entropy' (MaxEnt) comes from the fact that one would like the distributions to be as uniform as possible (at maximum entropy), by then introducing only the constraints or features that help to reduce the entropy to the level that resembles the actual class distribution in the data (Manning and Klein [2003]). The main advantage of MaxEnt models is that they can learn the most useful feature associations through feature weighting and inter-dependence analysis (Manning and Klein [2003], Wellner et al. [2006]), unlike the above-mentioned models which consider each feature to be independent of the others (Zaki and Meira [2010]). In addition, the output of MaxEnt models is easily interpretable, as features and classes are assigned a probability value that indicates the confidence of the classifier in its decision. This allows, e.g. via feature set analysis, to identify cases that are most difficult to classify, i.e. where the classifier has output low probability values on the classes and/or features.

As we have shown in Section 3.1.2, in previous work on connective disambiguation and especially when focusing on difficult types, the maximum entropy algorithm outperformed other ones. We have performed an empirical comparison over three connectives (*although*, *(even) though* and *since*) for SVM vs. MaxEnt classifiers, which is reported below in Section 5.5. This comparison showed that, over 26 feature subsets, in two thirds of the cases, the MaxEnt classifier outperformed the SVM one. The *a priori* and empirical arguments made us select the MaxEnt classifier for the experiments presented in this chapter. In fact, as observed also on other NLP problems, the performance of connective disambiguation appears to depend more strongly on the sets of features and classes (discourse relations or senses) than on the specific machine learning models that are employed.

5.2 Connective labeling vs. word sense disambiguation

The disambiguation of discourse connectives could be referred to as an instance of word sense disambiguation (WSD) where the task is to find meanings of words in context, e.g. the financial sense of the word *bank* vs. the river *bank*. WSD is generally applied only to content words (nouns, adjectives, verbs) rather than taking into account function words such as connectives.

The most obvious difference between WSD and connective labeling is that WSD concerns potentially all content words from a sentence, while connectives are sparse function words.

Insights from linguistics indicate that modeling the semantic meaning of content words differs considerably from modeling the procedural meaning of function words. The features needed to perform automatic WSD are quite different from those needed for connectives. Many WSD methods rely on local criteria, or sometimes on text-level topic models, which are not appropriate as features for discourse connectives, which require longer-range context features, as we show in the following.

We provide here a brief empirical argument demonstrating the need for connective-specific syntactic and semantic features. We implemented a baseline WSD system using as features only the two words preceding the occurrence of a discourse connective, and the three following ones. The system thus learns the word senses – here, the discourse relation labels – from a context window of five words, often considered sufficient for acceptable WSD performance. We used the SENSELEARNER system (Mihalcea and Csomai [2005]) to define models for the targeted word types and lists of senses, and experimented with it on our training data for the connective *while*, which has the most senses (five) and is the most difficult to classify (see Section 5.6.4). The training set for *while* consists of 236 occurrences from Europarl and 744 from PDTB, hence 980 occurrences, see Table 5.8). With 10-fold cross-validation on this set, SENSELEARNER reaches an average F1 score of 0.39.

Similarly, we built a Conditional Random Field (CRF) classifier (Lafferty et al. [2001]) which learned to label *while* with our sense labels, using as features the two words preceding each occurrence and their POS tags. With 10-fold cross-validation over the same training set, the F1 score was 0.47. Both scores are clearly lower than those obtained with the higher-level features we propose below, which are between 0.76 and 0.79 ($\pm$0.04) for 10-fold cross-validation experiments over the same data set. Therefore, typical WSD features do not appear to help much for the disambiguation of discourse connectives.

5.3 Features for connective labeling

Feature extraction Depending on our experimental settings, i.e. whether PDTB data and/or Europarl data have been used, the methods and tools for feature extraction vary. As described in Chapter 4, the PDTB annotation is an additional layer onto the Penn Treebank, a manual, gold-standard annotation of syntactical trees. All the PDTB files are easily linkable to PTB ones and therefore, no syntactical parser or POS tagger is needed to compute syntactical categories for connective features.

Additionally, because the two arguments of a connective are annotated as well in the PDTB, context word features can be extracted from these arguments directly. This is no longer the case when classifiers are trained on our own Europarl connective annotation, as no gold syntactic trees are available. We therefore made use of Charniak and Johnson [2005]'s syntactical parser to find the syntactic features. These features are noisier and more prone to errors, as the parser's performance is not fully accurate. Also, we cannot easily identify the arguments of a connective (automatic methods have been proposed, e.g. by Elwell and Baldridge [2008]

or Wellner and Pustejovsky [2007], but both with rather low precision). We therefore resort to use context words preceding and following the connective or appearing at the sentence boundaries.

The features used for discourse connective disambiguation include word-level and syntactic features already used in the past, as well as a series of novel semantically-oriented features. We will illustrate these features on an excerpt from the PDTB development set (WSJ_2448) with the connective *while* signaling CONTRAST:

> Hong Kong trade figures illustrate the toy makers' reliance on factories across the border. In 1989's first seven months, domestic exports fell 29%, to HK$3.87 billion, *while* re-exports rose 56%, to HK$11.28 billion.

The features are computed for the sentence containing the connective and for the preceding one (when available), thus accounting for possible inter-sentential dependencies and trying to get to similar features as when gold annotation for the arguments of a connective would be available, as especially argument 1 is often located in the previous sentence(s) (at least for connectives that can be coordinating conjunctions (e.g. *however, although*). For subordinating conjunctions (e.g. *while, since*) arguments 1 and 2 usually are located in the same sentence. Still, features from the preceding sentence are useful for wider context.

1. Surface features: words, POS, syntax and punctuation Previous studies (see 3.1) have reached above-random disambiguation scores by using surface features such as the connective word form (with the original capitalization), POS tags, and syntactic patterns from the hand-annotated parses provided by the Penn Treebank over the WSJ corpus. We therefore also use these features, obtaining them either from a re-ranking parser (Charniak and Johnson [2005]) or the syntactic trees available with the PTB. We extract a total of 9 word forms and 9 POS tags for each connective instance: the connective itself (with the capitalization of its first letter indicating the sentence-initial position), the words preceding and following it, as well as the words at the beginning and at the end of the sentence containing the connective and of the previous one. When only PDTB data is used, we replace the latter two context words by the ones at the beginning and end of the arguments of the connective. These context words often contain other connectives or connective-like expressions that can point to the sense of the connective to be found: *at the same time, but, by year end, when, and, if, etc.* The verb following the connective and the first verb in its sentence are also extracted from the parse trees. All word forms are lowercased after extraction, except the connective. For the example above, we obtain the following words and POS tags: *hong kong, NNP, border, NN, while, IN, billion, NN, re-exports, NNS, in, IN, billion, NN, fell, VBD, rose, VBD*. We also use as a feature the path of syntactic ancestors leading from the top of the parse tree to the connective, for which we build a pattern, e.g. $|S1||S||PP|$. Punctuation serves as another feature, which is encoded, following (Haddow [2005]), as *A.A,CA.* for the example sentences above, where *C* refers to the connective and *A* to all other words (i.e. there is the previous sentence up to the

period, followed by the beginning of the second sentence, a comma, the connective, followed by the end of the sentence).

2. Dependency features Dependency features can provide further indication on the syntactic role of connective (as coordinanting or subordinating conjunctions or adverbials) and their relations to other words in the sentences. We thus consider as another feature the dependency tags for the same 9 words as for the syntactic features above, using the output of Henderson's et al. dependency parser (Henderson et al. [2008]), along with the word position in the sentence. For the example above, the values are: *NAME, 1, ROOT, 14, TMP, 13, PMOD, 12, SBJ, 14, PMOD, 19, ROOT, SUB, 15.*

3. Auxiliary verbs In early work on automatic disambiguation of discourse connectives, Miltsakaki et al. [2005] have shown the usefulness of auxiliary verb features. Charniak and Johnson's parser tags them as *AUX*, which allows the extraction of *have, be, do* and *need* as auxiliary verbs. We generalize the auxiliaries in the same vein as Miltsakaki et al. [2005], with feature values of the form *AuxVerb_Tense* (with the auxiliary in its infinitive form) for all auxiliaries except when conjugated in present tense and third person singular, where the feature value e.g. becomes *has_third*. When no auxiliary verbs appear, as in the above example, the features remain unspecified.

4. WordNet features We attempt to detect pairs of words that are semantically related in the neighborhood of the connective. We extract from the parse tree the words before and after the connective, the first and last word of the sentence, the first verb in the sentence, and the first verb after the connective. We then compute lexical similarity scores for all 15 pairs of these six words using the Lesk metric (Banerjee and Pedersen [2002]), which measures the distance between two words in WordNet (Miller [1995]). The sum of these values is the value of the feature (0.10 in the above example). WordNet also provides semantic relations between lexical instances such as synonymy, meronymy and antonymy. The latter is especially relevant for our task, as we focus on connectives that frequently signal CONTRAST and CONCESSION. For the six words for which we compute the similarity scores, we query existing antonyms in WordNet. We then check in turn if one of those antonyms is present on the previous and current sentence, respectively. The feature value is the pair of actual antonyms found, e.g. in our example sentence: *fall-rise*. If no antonyms occur in the clauses, the feature remains unspecified.

5. TimeML features Some discourse connectives signal temporal relations (*meanwhile, since, while* and *yet*), which is why information on the temporal ordering of events is potentially helpful to detect those relations. We use the TimeML labels of temporal expressions as features, assigned automatically by the Tarsqi toolkit (Verhagen and Pustejovsky [2008]).

From the automatically annotated TimeML instances, we extract the main events in the sentence containing the connective and the preceding one, with their ordering and information on verb tenses and aspects. The value of this feature for the above example is the pattern *OCCURRENCE-PRES_OCCURRENCE-PAST* indicating an event in the present in the first sentence, and another event in the past in the second one.

6. Polarity features CONTRAST and CONCESSION, which can be signaled by *although, (even) though, however, while* or *yet*, are often accompanied by polar expressions such as negations or polar adjectives, verbs and nouns (e.g. *good, bad, increase, decrease, abuse* or *admiration*). To detect these expressions, we use a lexicon providing hand-annotated positive and negative sentiment values for about 8500 words (Wilson et al. [2005]). We look up all the words from the sentence containing the connective, and find their polarity value (e.g. 'negative_weaksubjective'). We then check, for each word, whether its five preceding words include negations and/or intensifiers (from a small hand-made list), and based on these elements we invert or, respectively, reinforce the polarity value obtained from the lexicon. Finally, we count the positive and negative polarity values for the text span preceding the connective, and the text span following it (until the end of the sentence), and generate four numeric feature values representing polarity. Moreover, we perform the same procedure for the preceding sentence, adding a fifth feature. For the above example, there is only one weak-subjective, negative word: *fell* (because *rose* is not in the polarity lexicon), resulting in the following values: *0, 0, 1, 0, 0*.

7. Discourse features The discourse connective labeling task can be seen as preliminary to discourse parsing, but this view can also be reversed. We use the output of the discourse parser by Soricut and Marcu [2003] as features for our connective labeler. Of course, if such a parser was fully accurate, it would *de facto* solve our task: however, this is far from being the case. The parser outputs a tree-like structure, where the nodes between text spans are labeled with one of the 128 RST discourse relations, which are informative for our task. The discourse feature consists of the concatenation of RST labels: one for the preceding sentence, one for the span of text preceding the connective and one for the span following it until the end of the sentence. For the example sentences the pattern is *Root. Joint-Joint, Contrast*, indicating that there is no discourse relation in the first sentence ('Root'), then the first span of the second sentence ('Joint') is in a paratactic relation with the second one ('Joint'), which contains a hypotactic relation of the type 'Contrast' starting at *while*.

8. Translational features The disambiguation model for discourse connectives is intended for MT systems. However, it can also benefit from the output of a baseline MT system, by using the hypothesized translation of a connective as an additional feature. Indeed, some occurrences of connectives may be translated by a connective that disambiguates them (e.g. *since* translated as *depuis que* for a TEMPORAL sense), correctly found by the MT system based on local constraints. We translate each discourse connective with a baseline Moses SMT

system from English into each target language for which the labeler will be combined with an MT system. The outputs are then realigned to the English source using Giza++. For all languages, the candidate translation, its position in the target sentence and its sense are the three features. The possible connective senses are inferred from a hand-made dictionary whose sense levels can be different granularity and precision. The first experiments with this feature used a simplified dictionary of connective senses (Meyer and Popescu-Belis [2012]) whereas in a later stage we relied on the more detailed dictionaries of the ACT translation metric (Hajlaoui and Popescu-Belis [2013] and Section 4.3) that in addition includes word alignment correction and the consideration of connective synonyms which results in more precise feature values. For the example sentence above, the French baseline translation provides the values *tandis que, 25, contrast*. These features are of course noisy: the baseline SMT contains errors (which our entire method aims to correct), the alignment is imperfect, and the baseline translation might not be specific enough.

Features for French connectives For the few preliminary experiments on the disambiguation of French connectives, the features slightly differ from the English ones, as less sophisticated NLP tools are available. The French features were the following: the sentence-initial character of the connective (yes/no); the dependency tag of the connective; the first verb in the sentence; its dependency tag; the word preceding the connective; its POS tag; its dependency tag; the word following the connective; its POS tag; its dependency tag; the first verb after the connective; and its dependency tag. The French texts were POS-tagged with the MElt tagger (Denis and Sagot [2009]) and parsed with the MaltParser (Nivre [2003]), which generates dependency trees. In contrast to constituents, dependency structures contain information about the grammatical function of each word (heads) and link the dependents belonging to the same head. However, as the dependency parser provides no differentiated verb tags (as auxiliaries), we extracted the verb word forms and added their dependency tags. The same applies to the connective itself, and preceding and following words and their dependency tags. The dependency tag of the non-connectives varies between *subj* (subject), *det* (determiner), *mod* (modifier) and *obj* (object). The first verb in the sentence often belongs to the *root* dependency while the verb following the connective most often belongs to the *obj* dependency. For *alors que*, the most frequent dependency tags were *mod_mod* and *mod_obj*, indicating the connective's main function as a modifier of its argument.

5.4 Disambiguation experiments based on the PDTB

Our first experiment was aimed at sense disambiguation down to the third level of the PDTB hierarchy. We used the WEKA machine learning toolkit (Hall et al. [2009]) and its implementation of a Random Forest classifier (Breiman [2001]). This method outperformed, in our task, the C4.5 decision tree and Naive Bayes algorithms sometimes used in research on discourse connective classification. The training set here consisted of all 100 types of explicit connectives annotated in the PDTB training set (15,366 instances). To make the figures and results compa-

Connective	Senses with number of occurrences	Accuracy	Baseline	*kappa*
although	134 CO, 133 CT	**58.4%**	48.7%	0.17
but	2090 CT, 485 CO, 77 E	76.4%	78.8%	0.02
however	261 CT, 119 CO	68.4%	68.7%	0.05
meanwhile	77 T, 57 E, 22 CT	51.9%	49.4%	0.09
since	83 C, 67 T	**75.3%**	55.3%	0.49
though	136 CO, 125 CT	**65.1%**	52.1%	0.30
when	640 T, 135 COND, 17 C, 8 CO, 2 CT	79.9%	79.8%	0.05
while	342 CT, 159 T, 77 CO, 53 E	**59.6%**	54.1%	0.23
all conn.	2975 CT, 959 CO, 943 T, 187 E, 135 COND, 100 C	**72.6%**	56.1%	0.50

Table 5.1: Accuracy for the disambiguation of eight English temporal–contrastive connectives with a Random Forest classifier. The connective senses are encoded as follows: CO: CONCESSION, CT: CONTRAST, E: EXPANSION, T: TEMPORAL, COND: CONDITION, and C: CAUSE. The last line (*all conn.*) provides the results of an independent classification experiment with the eight connective types and six classes – it is not the average over the eight classifiers specific to each connective.

rable to related work, we used the subdivision of the PDTB recommended in the annotation manual (Prasad et al. [2007]): sections 02–21 as training set and section 23 as test set. The only two features were the (capitalized) connective word tokens from the PDTB and their Part of Speech (POS) tags. For *all 129 possible sense combinations*, including complex senses, results reach *66.51% accuracy* with 10-fold cross validation on the training set and *74.53% accuracy* on the PDTB test set [1]. This can be seen as a baseline experiment. Another baseline was reported by Pitler and Nenkova [2009] with accuracy of 85.86% for correctly classified connectives (with the 4 main senses), when using the connective token as the only feature.

Based on an analysis of translations and frequencies, we then reduced the list of PDTB senses (Figure 4.1) to the following six: TEMPORAL (T), CAUSE (C), CONDITION (COND), CONTRAST (CT), CONCESSION (CO) and EXPANSION (E). All subsenses from the third PDTB hierarchy level were merged under second level ones (C, COND, CT, CO). Exceptions were the top level senses T and E, which, so far, need no further disambiguation for translation. In addition, we extracted separate training sets for each of the 8 connectives *although, but, however, meanwhile, since, though, when* and *while*. The number of occurrences and senses in the sets for the single connectives is listed in Table 5.1. The total number of instances in the training set for all 8 connectives is 5,299 occurrences, with a sense distribution of 56.1% CT, 18% CO, 17.8% T, 3.5% E, 2.5% COND, 1.9% C. The features extracted from the PDTB were the ones described above in Section 5.3, group 1.

Results were generated separately for every temporal–contrastive connective (assuming the goal is to improve the translation of only certain connectives), in addition to one result for

1. As far as we know, Versley [2010] is the only reference reporting results down to the third level, reaching an accuracy of 79%, using more features, but not stating whether the complex sense annotations were included.

the entire subset. The results in Table 5.1 above are based on 10-fold cross validation on the training sets. They were measured using accuracy (percentage of correctly classified instances) and the $kappa$ value. The baseline is the majority class, i.e. the prediction for the most frequent sense annotated for the corresponding connective. Marked in bold are the accuracy values significantly above the baseline ones[2]. In the experiment with global classification for all eight temporal–contrastive connectives and all six sense classes (last line of Table 5.1), the accuracy and $kappa$ values are well above random agreement or the prediction of the majority class.

Experiments for specific subsets of connectives have rarely been reported in the literature. Miltsakaki et al. [2005] describe results for *since, while* and *when*, reporting accuracies of 89.5%, 71.8% and 61.6%. The results for the single connectives are comparable with ours in the case of *since* and *while*, where similar senses were used. For *when* they only distinguished three senses, whereas we report a higher accuracy for 5 different senses, shown in Table 5.1. We provide elsewhere (Meyer [2011]) more details on our experiments.

These initial experiments and results confirmed that the temporal–contrastive connectives that are problematic in translation can automatically be disambiguated with state of the art performance. We however encountered memory problems with the Random Forest classifier when wanting to add more features and/or more detailed sense labels. Also, to compare further to other state of the art methods, we use a maximum entropy algorithm in a further experiment with PDTB data, more types of connectives, more features and more elaborate sets of senses. A subset of 13 ambiguous, again mainly temporal–contrastive connectives was the training material, selected on previous corpus studies that identified these connectives as being especially problematic and ambiguous for translation. For each connective we built a specialized classifier and extracted more features then before: to the basic set of group 1 features we added the WordNet (group 4) and TimeML (group 5) features described in Section 5.3, because they help disambiguating temporal connectives (given TimeML information) and contrastive ones (antonym information from WordNet). The details on these classifiers are published in (Meyer and Popescu-Belis [2012]).

We report the classifier performances as micro-averaged F1 scores for each connective in Table 5.2, testing on Section 23 of the PDTB.

In an attempt to further improve these models, we added a new type of feature, namely the one using candidate translations of discourse connectives from a baseline SMT system (not adapted to connectives) as was mentioned for feature group 8 in Section 5.3, here in the version making use of simple sense dictionaries as at the time the ACT metric for connectives was not yet available. Overall, this procedure led to accuracy gains of about 0.1 to 0.6 F1 score for some of the connectives, as can be seen in the last column of Table 5.2. These scores are well above the ones from the preliminary experiment in Table 5.1 and sometimes also higher

2. Paired t-tests were performed at 95% confidence level. The other accuracy values are either near to the baseline ones or not significantly below them.

Connective	Number of occurrences and senses				F1 Scores	
	Train. set: total and per sense		Test set: total and per sense		PT	PT+
after	507	456 As, 51 As/Ca	25	22 As, 3 As/Ca	0.66	1.00
although	267	135 Cs, 118 Ct, 14 Cp	16	9 Ct, 7 Cs	0.60	0.66
however	176	121 Ct, 32 Cs, 23 Cp	14	13 Ct, 1 Cs	0.33	1.00
indeed	69	37 Cd, 24 R, 3 Ca, 3 E, 2 I	*2	2 R	*0.50	*0.50
meanwhile	117	66 Cj/S, 16 Cd, 16 S, 14 Ct/S, 5 Ct	10	5 S, 5 Ct/S	0.32	0.53
nevertheless	26	15 Ct, 11 Cs	6	4 Cs, 2 Ct	0.44	0.66
nonetheless	12	7 Cs, 3 Ct, 2 Cp	*1	1 Cs	*1.00	*1.00
rather	10	6 R, 2 Al, 1 Ca, 1 Ct	*1	1 Al	*0.00	*0.00
since	166	75 As, 83 Ca, 8 As/Ca	9	4 As, 3 Ca, 2 As/Ca	0.78	0.78
still	114	56 Cs, 51 Ct, 7 Cp	13	9 Ct, 4 Cs	0.60	0.66
then	145	136 As, 6 Cd, 3 As/Ca	6	5 As, 1 Cd	0.83	1.00
while	631	317 Ct, 140 S, 79 Cs, 41 Ct/S, 36 Cd, 18 Cp	37	19 Ct, 10 S, 4 Cs, 4 Ct/S	0.93	0.96
yet	80	46 Ct, 25 Cs, 9 Cp	*2	2 Ct	*0.5	*1.00
Total	**2,320**	–	**142**	–	**0.57**	**0.75**

Table 5.2: Performance of MaxEnt connective sense classifiers: *Classifier PT* (feature groups 1, 4 and 5) and *Classifier PT+* (with features from group 8) for 13 temporal and contrastive connectives in the PDTB. The sense labels here are named as the ones in the PDTB (Figure 4.1), from either the first or the second level of the sense hierarchy: Al: alternative, As: asynchronous, Ca: cause, Cd: condition, Cj: conjunction, Cp: comparison, Cs: concession, Ct: contrast, E: expansion, I: instantiation, R: restatement, S: synchrony. In some cases marked with '*', the test sets are too small to provide meaningful scores.

than the state-of-the-art (Versley [2011]), which is the only study that a) built single classifiers for many different connectives and b) used a set of fine-grained relations from the second level of the PDTB hierarchy (although not reporting whether the double senses were included, which we did). The translational features often help to outperform the state of the art (Versley [2011]) for *nevertheless* (0.53 vs. 0.66), *although* (0.61 vs. 0.66), *still* (0.51 vs. 0.66), *while* (0.72 vs. 0.96), *yet 0.65 vs. 1.00*. In other cases our classifier performance is worse than (Versley [2011]): *rather* (0.64 vs. 0.00), *since* (0.93 vs. 0.78), *meanwhile* (0.86 vs. 0.53). Versley [2011] however trained and tested on occurrences from the PDTB training set (sections 2-22), with cross-validation which is why the scores are only indirectly comparable with ours. We will repeat this comparison when reporting our cross-validation experiments on the training set in Section 5.6.

On the one hand, the classifiers trained on the PDTB could directly be applied to SMT by tagging a subsection of the Europarl corpus that is the training material for SMT. On the other hand however, there is a genre and register change involved from the newswire texts of the WSJ corpus to formal, political speech in Europarl. Moreover, certain annotation errors by

Connective	Labels	Baseline	R. Forest		N. Bayes		SVM	
		Acc.	*Acc.*	κ	*Acc.*	κ	*Acc.*	κ
alors que	B, C, B/C	46.9	*53.1*	*0.2*	***55.7***	***0.3***	*54.2*	***0.3***
alors que	B, C	68.7	**69.2**	0.1	68.3	***0.2***	64.7	0.1
since	T, CA, T/CA	51.6	*79.8*	*0.6*	*82.3*	***0.7***	***85.4***	***0.7***
since	T, CA	51.6	*80.7*	*0.6*	*84.0*	***0.7***	***85.7***	***0.7***
while	T/C, T/PUNCT, T/DUR, T/CA, CONC, C	44.8	43.2	*0.1*	49.9	***0.2***	***52.2***	***0.2***
while	T, C, CONC	43.5	*60.5*	***0.3***	*59.9*	***0.3***	***60.9***	***0.3***

Table 5.3: Disambiguation scores for three connectives with two sets of labels each, for various classification algorithms. Accuracy (*Acc.*) is in percentage, and $kappa$ is always zero for the baseline method (majority class). The best scores for each data set are in **boldface**, and scores significantly above the baseline (95% t-test) are in *italics*. The sense labels are encoded as follows: b: BACKGROUND, c: CONTRAST, ca: CAUSAL, conc: CONCESSION, t: TEMPORAL, punct: PUNCTUAL, dur: DURATIVE.

classifiers trained on PDTB data will be propagated into the SMT process which in turn will lead to losses in performance for the translation task. To address this limitation, the following sections describe the automatic labeling of connectives in Europarl and joint Europarl and PDTB data.

5.5 Disambiguation experiments based on Europarl

Taking advantage of the annotations of discourse connectives in the Europarl corpus that we made available (see Chapter 4), we performed a series of experiments over several datasets (listed in Table 4.4), in order to test whether the sense labels obtained through translation spotting and clustering are useful for automatic classification.

A first series of classification experiments in English and French (Meyer et al. [2011]) made use of the WEKA machine learning toolkit (Hall et al. [2009]) to compare several classification algorithms: Random Forest, Naive Bayes, and Support Vector Machine. The results are reported with 10-fold cross validation on the entire dataset for each connective, using all features from group 1 (Section 5.3).

Table 5.3 lists for each method – including the majority class as a baseline – the percentage of correctly classified instances (or accuracy, noted *Acc.*), and the *kappa* values. Significance above the baseline is computed using paired t-tests at 95% confidence. When a score is significantly above the baseline, it is shown in *italics* in Table 5.3. The best scores for each dataset, across classifiers, are indicated in **boldface**. When these scores were not significantly above the baseline, at least they were never significantly below either.

In two cases, the SVM classifier performed best: the maximum accuracy for *since* is 85.7%,

for *while* it is 60.9%. For *alors que*, the maximum accuracy of 69.2% is reached with Random Forest.

The analysis of results for each data set leads to observations that are specific to each connective. The high improvement over the baseline for the first experiment on *alors que* confirms the usefulness of the double-sense B/C label for this connective and supports the idea that the temporal and the contrastive meanings may co-exist. Comparatively, the classifier using two labels only marginally and not significantly outperforms the baseline score of 68.7%. Although its absolute accuracy is much higher with respect to the three-way classifier (69.2% vs. 55.7%), its actual improvement with respect to the baseline (majority class) is very low, as correctly captured by the *kappa* score, which is higher for the three-way classifier. While more elaborate features may help, these scores can be related to the difficulties of human annotators in disambiguating *alors que* (see Chapter 4, Section 4.2.1).

R	**Feature**	**IG**	
		S1	S2
1	preceding word	1.12	0.64
2	following verb	0.81	0.51
3	first verb	0.74	0.42
4	following word	0.68	0.23
5	preceding word's POS tag	0.15	0.05
5	first verb's dep. tag	0.14	0.06
5	following word's POS tag	0.19	0.03
8	preceding word's dep. tag	0.10	0.03
8	connective's dep. tag	0.09	0.04
10	following word's dep. tag	0.13	0.013
10	following verb's dep. tag	0.04	0.03
12	sentence initial	0.05	0.001

Table 5.4: Information gain (IG) of features for French connective *alors que*, ordered by decreasing average ranking (R) in both sense settings (S1 and S2). Features 1–4 are considerably more relevant than the following ones.

The features used so far lead to high scores for *since* in both datasets. The SVM classifier outperforms considerably the one used by Miltsakaki et al. [2005] on the three-way classification task (with T, C, T/CA), with an accuracy of 85.4% vs. 75.5%, obtained however on different datasets. For the two-way classification (T, CA), again on different datasets, our accuracy of 85.7% is slightly lower than the 89.5% given in (Miltsakaki et al. [2005]).

For *while*, when comparing the first set of senses against the second one, it appears that reducing the number of labels from six to three increases accuracy by 8-10%. This is due to the small number of training instances for the labels T/PUNCT and T/DUR in the first setting. However, even for the larger set of labels, the scores are significantly above baseline (52.2% vs. 44.8%), which indicates that such a classifier can still be useful as input to an MT system,

possibly improved thanks to a larger training set. The performance obtained by Miltsakaki et al. [2005] on *while* is markedly better than ours, with an accuracy of 71.8% compared to ours of 60.9% with three labels.

When comparing the scores on Europarl data with the ones that were given in the previous section on the PDTB, the scores on Europarl data were slightly higher for *since* (85.7% vs. 78%) and much lower for *while* (60.9% vs. 96%). This is due to variation in features, label sets, training/test data sizes and the usage of different classification algorithms all of which will be consolidated in experiments that use both datasets in Section 5.6.

R	**Feature**	**IG**	
		S1	S2
1	preceding word	0.83	0.75
2	following word	0.56	0.52
3	following verb's POS tag	0.24	0.21
4	type of following aux. verb	0.13	0.12
5	type of first aux. verb	0.11	0.11
6	first verb's POS tag	0.02	0.01
7	sentence initial	0.00	0.00

Table 5.5: Information gain (IG) of features for EN connective *since*, ordered by decreasing average ranking (R) in both experimental settings S1 and S2.

The relevance of features can be measured by computing the information gain (IG) brought by each feature to the classification task, i.e. the reduction in entropy with respect to desired classes (Hall et al. [2009]) – the higher the IG, the more relevant the feature. Features can be ranked by decreasing IG, as shown in Tables 5.4, 5.5 and 5.6, in which ranks were averaged over the first and the second data set in each series.

R	**Feature**	**IG**	
		S1	S2
1	preceding word	1.02	0.65
2	following word	0.83	0.55
3	type of first aux. verb	0.12	0.07
4	following verb's POS tag	0.16	0.04
5	first verb's POS tag	0.07	0.09
5	type of following aux. verb	0.12	0.05
7	sentence initial	0.08	0.07

Table 5.6: Information gain (IG) of features for EN connective *while*, ordered by decreasing average ranking (R) in experiments with sense settings S1 and S2. The first two features are considerably more relevant than the remaining ones.

The tables show that across all three connectives and the two languages, the contextual features are always in the first positions, thus confirming the importance of the context of a connective.

Following these are verbal features, which are, for these connectives, of importance because the temporal meanings are additionally established by verbal tenses. POS and dependency features seem the least helpful for disambiguation.

Connective	**Number of occurrences and senses**				**F1**
	Training set: total and per sense		Test set: total and per sense		**Score**
although	168	150 Cs, 18 Ct	15	10 Cs, 5 Ct	0.92
meanwhile	103	92 S, 11 Ct	28	25 S, 3 Ct	1.00
since	341	222 S, 111 Ca, 8 S/Ca	82	55 S, 25 Ca, 2 S/Ca	1.00
(even) though	277	202 Cs, 75 Ct	69	50 Cs, 19 Ct	1.00
while	237	108 Cs, 74 S/Ct, 35 Ct, 11 S/Ca, 9 S	57	26 Cs, 18 S/Ct, 8 Ct, 3 S/Ca, 2 S	0.73
yet	323	169 Adv, 106 Cs, 48 Ct	77	40 Adv, 25 Cs, 12 Ct	1.00
Total	**1449**	–	**328**	–	**0.94**

Table 5.7: Training/test data and performance (macro-average F1 scores) of the automatic connective sense labeler, for seven highly-ambiguous connectives annotated over the Europarl Corpus. The sense labels are coded as follows. Cs: Concession, Ct: Contrast, S: Synchrony, Ca: Cause, Prep: Preposition, Adv: Adverb.

In further experiments (Meyer et al. [2012]) we started to use the larger feature sets (feature groups 1, 4 and 5), with maximum entropy models for classification. For the six connectives *although, meanwhile, since, (even) though, yet*, in Table 5.7, we report the results of these classifiers, again in terms of F1 scores. Using maximum entropy and more features now clearly put the scores in the very same range as they were for the PDTB data (Section 5.4).

Still, the test sets of these 'Europarl-classifiers' are very small, and from the fact that SMT will deal with Europarl and newswire data (the latter for tuning and test), the classifiers will best be implemented not only with more training data, but specifically with a combination of Europarl *and* PDTB data, which will be described in the next section, along with a detailed evaluation of 10-fold cross-validation, feature analysis and results on different test sets.

5.6 Experiments on large feature and data sets

The methods and results of this section are further discussed in (Meyer et al. [2014]).

5.6.1 Merging PDTB and Europarl data

One could think of using the existing PDTB gold-standard annotations directly, either for training disambiguation modules (as was shown above) or directly for SMT. This however has the considerable disadvantage that there is no human translation of the WSJ corpus from English into another language (except for Czech, see Section 7.1.1), which means that

Connective	Training set			Testing set		
	EP	PDTB	Distribution of labels (%)	EP	PDTB	Distribution of labels (%)
although	168	312	Ct: 68.9; Cs: 31.1	15	16	Ct: 48.4; Cs: 51.6
however	348	450	Ct: 47.8; Cs: 52.2	70	35	Ct: 47.6; Cs: 52.4
meanwhile	102	177	Ct: 77.3; T: 22.7	28	14	Ct: 76.2; T: 23.8
since	339	174	Ca: 38.7; T: 59.6; T/Ca: 1.7	82	10	Ca: 30.4; T: 67.4; T/Ca: 2.2
(even) though	276	306	Ct: 33.3; Cs: 66.7	69	14	Ct: 33.7; Cs: 66.3
while	236	744	Ct: 14; Cs: 23; T: 15; T/Ct: 46.6; T/Ca: 1.4	58	37	Ct: 22.8; Cs: 33.7; T: 9.8; T/Ct: 30.4; T/Ca: 3.3
yet	326	99	Ct: 23.2; Cs: 29.8; Adv: 47	77	2	Ct: 30.4; Cs: 19; Adv: 50.6
Total	1795	2262	–	399	128	–

Table 5.8: Numbers of connectives and distributions of labels in the training and test sets for connective labeling, from Europarl (EP) and the Penn Discourse Treebank (PDTB). Ct: CONTRAST, Cs: CONCESSION, T: TEMPORAL, Ca: CAUSAL, Adv: ADVERB.

there is no reference translation to which a discourse-aware MT system could be compared. Furthermore, SMT training is often performed on the Europarl corpus, whereas the WSJ corpus consists of newswire text. This shift in topic and genre affects automated classification performance when only training on one of the two genres.

To offer a larger amount of training data for the classification task, we merged the Europarl and the PDTB annotated datasets. For each of the the seven discourse connectives *although, however, meanwhile, since, (even) though, while, yet*, we first extracted all the explicit instances from the PDTB in accordance to the recommendation given in the PDTB manual (Prasad et al. [2007]), i.e. using WSJ Sections 02-21 for training, Sections 00, 01, 22, and 24 for development and Section 23 for testing. Then, we split the Europarl dataset (Table 4.4) into a training and a test part, as can be seen in Table 5.8 [3]. To merge these sets with the PDTB ones, we mapped the PDTB senses ([Prasad et al., 2007, p. 27]) to those we defined for Europarl, using the following rules:

— *although, (even) though, however:* if one of the PDTB labels is EXPECTATION or CONTRA-EXPECTATION, then convert the label to CONCESSION; otherwise to CONTRAST.
— *since:* if it is labeled CONTINGENCY and TEMPORAL, then convert the label to TEMPORAL-CAUSAL (composite label); if it is only labeled CONTINGENCY, then convert it to CAUSAL; otherwise to TEMPORAL.
— *meanwhile:* if one of the PDTB labels is COMPARISON, then convert the label to CON-

3. The Europarl training and test sets (without the PDTB parts) are freely available at: https://www.idiap.ch/dataset/Disco-Annotation.

Features	although	however	meanwhile	since	though	while	yet
(Majority class)	0.69	0.52	0.77	0.60	0.67	0.47	0.47
Sentence_initial	0.49	0.60	0.81	0.57	0.49	0.52	0.74
Words	0.72	**0.88**	**0.85**	0.91	**0.76**	0.77	**0.90**
POS_tags	0.65	0.73	0.82	0.76	0.70	0.57	0.81
Punctuation	0.49	0.30	0.81	0.66	0.70	0.60	0.73
Syntax	0.57	0.62	0.78	0.61	0.52	0.61	0.53
All_Syntactic	**0.75**	0.85	**0.85**	**0.96**	**0.76**	**0.78**	0.87
Dependency	**0.69**	**0.82**	**0.88**	**0.90**	**0.80**	**0.73**	**0.83**
WordNet	0.55	0.73	0.81	0.69	0.61	0.58	0.46
Auxiliary_Verbs	0.52	0.63	0.74	0.72	0.54	0.51	0.43
TimeML	0.58	0.70	0.81	0.60	0.55	0.58	0.49
Translational	0.49	0.64	0.81	0.71	0.63	0.53	0.75
Polarity	0.48	0.63	0.82	0.64	0.49	0.48	0.35
Discourse	0.51	0.56	0.78	0.69	0.56	0.52	0.37

Table 5.9: F1 scores for connective labeling (10-fold c.-v.) for each type of syntactic and semantic features. The best scores per connective for each of the two types are **in bold**. *Though* also includes occurrences of *even though* (and considered as connective of two words, i.e. the preceding word is not *even* but another word).

TRAST; otherwise to TEMPORAL.

— *while:* if it is labeled COMPARISON and TEMPORAL, then convert the label to TEMPORAL-CONTRAST (composite label); if it is labeled TEMPORAL and another label (different from COMPARISON), then convert it to TEMPORAL; if it is labeled EXPANSION, or PRAGMATIC-CONTRAST, or CONJUNCTION, then convert it to CONTRAST (closest common sense); otherwise to CONCESSION.

— *yet:* if it is labeled EXPANSION, then it can be considered to behave close to adverbial usage and is therefore labeled ADVERB; if one of the PDTB labels is EXPECTATION or CONTRA-EXPECTATION, then convert the label to CONCESSION; otherwise to CONTRAST.

While our labels tend to correspond to the PDTB's second level, we also consider labels encoding two senses, unlike previous work on automatic labeling which considers only the first (most general) sense.

5.6.2 Feature analysis and selection

To estimate the contribution of each feature, we started by testing them individually, using 10-fold cross-validation. Then, we grouped the surface and syntactic features (group 1 in Section 5.3 above) into a set called `All_Syntactic` and tested it as well. The results of these experiments are shown in Table 5.9.

The `All_Syntactic` set appeared to outperform all other features considered individually, including the semantic ones, echoing previous results by Pitler and Nenkova [2009]. Still, the

Dependency features (group 2, Section 5.3), which are the best performing semantic features, are close to All_Syntactic, and even outperform them for two connectives (*meanwhile* and *(even) though*).

Feature subsets	although	however	meanwhile	since	though	while	yet
All_Synt+Dependency	0.73	0.85	0.85	0.93	0.76	0.78	**0.90**
All_Synt+WordNet	0.73	0.85	0.83	**0.96**	0.75	0.78	0.87
All_Synt+Auxiliary_Verbs	0.74	**0.87**	0.83	0.94	0.76	0.77	**0.90**
All_Synt+TimeML	0.72	0.86	0.86	0.92	0.73	**0.79**	0.87
All_Synt+Translational	**0.75**	0.87	0.85	0.91	**0.77**	0.77	**0.90**
All_Synt+Polarity	0.74	0.86	**0.87**	0.95	0.74	0.78	0.89
All_Synt+Discourse	0.72	0.86	0.83	0.95	0.76	0.78	0.88
All_Synt+Dep+Trans	**0.71**	0.85	0.85	**0.93**	0.77	0.77	**0.90**
All_Synt+Dep+Trans+TimeML	0.70	**0.86**	0.86	0.93	**0.78**	**0.78**	0.90
All_Synt+Dep+Trans+TimeML+WN	0.71	0.86	0.86	0.92	0.78	0.77	0.90
All_Synt+Dep+Trans+TimeML+WN+Aux	0.71	0.86	0.85	0.91	0.78	0.77	0.90
All_Synt+Dep+Trans+TimeML+WN+Aux+Disc	0.70	0.85	**0.87**	0.91	0.77	0.76	0.89
All_Features	0.69	0.85	0.86	0.93	0.77	0.76	0.88

Table 5.10: F1 scores for connective labeling (10-fold c.v.) for combinations of features, always including all syntactic features (All_Synt) and in the lower half the dependency ones (Dep). The best scores per connective and group are **in bold**. *Though* also includes occurrences of *even though*.

A second series of tests, shown in the upper half of Table 5.10, was performed by using for classification the All_Syntactic subset of features, plus each of the semantic features separately (7 experiments). Then, a third series of tests (lower half of Table 5.10) was performed by incrementing gradually the feature set, from All_Syntactic, with the semantic features ordered by decreasing average of individual performance, as indicated in the table. Finally, the last line of Table 5.10 provides the scores of the All_Features model.

From these experiments, it appears that performance increases quite modestly when adding more features. The variations for each connective, especially in the lower half of Table 5.10, are quite small. The highest scores for each connective are reached with different subsets, and the best scores for All_Syntactic plus the best-performing semantic feature are generally slightly higher than those for All_Features.

Classification scores close to the best ones can be reached by using the surface and syntactic features only, as found also in previous work (Pitler et al. [2008], Pitler and Nenkova [2009]). However, the All_Syntactic models are always outperformed when adding features from the dependency parses. Moreover, the Dependency and All_Syntactic + Dependency models for each connective reached particularly high scores. Therefore, using All_Syntactic + Dependency models appears to be a recommendable strategy, which is applicable to a larger range of languages than the models with the higher-level semantic features. However, overall, it is best to use the complete feature set we defined because of its robustness on the test sets (as we will conclude in Section 5.6.4).

In any case, a separate classifier should be used for each discourse connective. We tested, with

Features	although	however	meanwhile	since	though	while	yet
Sentence_initial	−	−	−	−	−	−	−
Words	+	·	−	·	·	·	·
POS_tags	·	−	−	−	−	−	·
Punctuation	−	−	−	−	−	−	−
Syntax	−	−	−	−	−	−	−
All_Syntactic	+	·	−	·	·	·	·
Dependency	·	·	·	·	·	·	·
WordNet	−	−	−	−	−	−	−
Aux	−	−	−	−	−	−	−
TimeML	−	−	−	−	−	−	−
TR	−	−	−	−	−	−	−
Polarity	−	−	−	−	−	−	−
Discourse	−	−	−	−	−	−	−
All_Synt+Dep	+	·	·	·	·	·	·
All_Synt+WN	+	·	−	·	·	·	·
All_Synt+Aux	+	·	−	·	·	·	·
All_Synt+TimeML	·	·	·	·	−	·	·
All_Synt+Trans	+	·	−	·	·	·	·
All_Synt+Pol	+	·	·	·	·	·	·
All_Synt+Disc	·	·	−	·	·	·	·
All_Synt+Dep+Trans	·	·	·	·	·	·	·
All_Synt+Dep+Trans+TimeML	·	·	·	·	·	·	·
All_Synt+Dep+Trans+TimeML+WN	·	·	·	·	·	·	·
All_Synt+Dep+Trans+TimeML+WN+Aux	·	·	−	·	·	·	·
All_Synt+Dep+Trans+TimeML+WN+Aux+Disc	·	·	·	·	·	·	·

Table 5.11: Comparison of the F1 score of each feature subset against All_Features used for connective labeling. Significant improvements (10-fold c.v., 95% level) are noted with +, significant degradations with −, and the absence of significant differences is noted ·. Overall, using All_Features is never outperformed by any subset, except for *although*. *Though* also includes occurrences of *even though*.

10-fold cross-validation, a unique classification model for all seven discourse connectives with all features. This model reached 0.80 F1 score, which is only slightly, but significantly, lower than when averaging over the seven single connective classifiers with All_Features, which results in 0.82 F1 score. This corroborates previous results on comparing item-specific vs. joint classifiers for discourse markers, (e.g. Popescu-Belis and Zufferey [2007], Versley [2011]).

5.6.3 Significance of connective labeling scores

In Table 5.11, we provide an assessment of the statistical significance of the differences in scores, with respect to the All_Features model, of the various feature subsets used for connective labeling listed in the first columns of Tables 5.9 and 5.10. To note the result of the significance test, when a subset of features performs significantly better, at 95% confidence using 10-fold c.v., than the All_Features model, this is indicated by a '+' sign. Conversely, significantly lower performance is indicated with a '−', and no significant difference is indicated by with a '·' sign. Overall, it can be observed that there are only a few cases where a feature subset significantly outperformed the All_Features model, all related to *although*, as shown by the '+' signs in Table 5.11.

Data	Method	although	however	meanwhile	since	though	while	yet
Training (c.v.)	AF	0.69±0.04	0.85±0.05	0.86±0.01	0.93±0.05	0.77±0.04	0.76±0.04	0.88±0.07
Test: Europarl and PDTB (WSJ s. 23)	MC	0.52	0.52	0.76	0.68	0.66	0.34	0.51
	AF	0.58	**0.73**	0.71	**0.90**	0.69	0.45	**0.78**
	Best	0.61	0.60	0.74	0.87	**0.71**	0.43	0.72
	Synt+Dep	**0.65**	0.67	**0.79**	0.89	0.7	**0.47**	0.72
Test: Europarl	AF	0.60	**0.69**	0.79	**0.90**	0.67	0.45	**0.78**
	Best	**0.80**	0.56	0.82	0.85	**0.72**	0.43	0.74
	Synt+Dep	0.73	0.66	**0.89**	0.88	0.71	**0.50**	0.73
Test: PDTB (WSJ s. 23)	AF	**0.56**	**0.83**	0.57	0.90	**0.79**	**0.46**	**1.0**
	Best	0.44	0.69	0.57	**1.0**	0.64	0.43	0.0
	Synt+Dep	**0.56**	0.69	0.57	**1.0**	0.64	0.43	0.50

Table 5.12: F1 score on test data for connective labeling with the `All_Features` (AF) model, with the best model found on the training data (`Best`), and with syntactic and dependency features only (`Synt+Dep`). The proportion of the majority class (MC) on the EP+PDTB test set is indicated as a baseline, along with the F1 score of `All_Features` on the training data, with confidence intervals. *Though* also includes occurrences of *even though*.

Most of the differences in scores are not significant. From our analysis, it appeared that there was only one connective, *although*, for which the `All_Features` model was significantly outperformed by certain feature subsets (e.g. `All_Synt + Polarity`). This smaller number of features here was sufficient, while the data size for *although* was not sufficient to learn a model using `All_Features`.

5.6.4 Results on the test sets

We tested the accuracy of our best classifiers found on the training data on three previously unseen sets: a test set from Europarl, another one from the PDTB, and their union noted EP+PTDB. We evaluated for each of the connectives and for each test set the best-scoring MaxEnt model (i.e. with the best feature set) found on the training data (noted `Best`), the `All_Syntactic + Dependency` model, and the `All_Features` model. The F1 scores are shown in Table 5.12, adding in the first line the performance of the `All_Features` model on the training data with 95% confidence intervals computed from the 10 folds. Almost all classifiers outperform significantly the scores of the majority class baselines (proportion of the largest class in Table 5.8). Only the classifiers for *meanwhile* sometimes perform below their baseline (due to the high majority class of 0.76), whereas substantial improvement is gained for all other classifiers, with *yet* outperforming its baseline the most (0.88±0.07 vs. 0.51). In terms of F1 scores on the combined Europarl+PDTB test set, the highest disambiguation performances are 0.90 for *since*, 0.79 for *meanwhile* and 0.78 for *yet*.

Moreover, the `All_Features` model scored best 11 times on the three test sets, versus 4 times for `Best` and 7 times for `Syntactic+Dependency`. Therefore, one best generates the complete feature set to tag the instances for translation appears to be the best and most general option. For the SMT systems we will have a similar mixture of Europarl and newswire data, which is why the use of all features can most reliably capture the properties of both text genres.

The scores confirm that very much of the performance can be gained by using syntactic features plus dependency ones, although the use of `All_Features` is the most reliable strategy. From both training and test set scores one can also see that *since* is the easiest connective to disambiguate, with F1 scores from 0.85 to 1.0. The connective *while* has reasonable training scores (around 0.76), but its ambiguity is hard to resolve on unseen test data where performance drops down to 0.43, although still above the baseline.

Our classifiers compare favorably to the state of the art for classifying highly-ambiguous connectives (reviewed in Section 3.1). We hypothesize that this is due to the specialized features we defined. Moreover, this is – to the best of our knowledge and besides our own previous work (Meyer [2011], Meyer et al. [2011]) – the first attempt to automatically disambiguate some of the composite senses of ambiguous connectives. Our pre-trained models for `All_Syntactic + Dependency` features and the feature extractors are publicly available [4].

	nt2008+sy2009		nt2010		nt2012	
Connective	*P* %	*F1*	*P* %	*F1*	*P* %	*F1*
although	16	0.60	4	0.57	9	0.63
however	35	0.53	26	0.65	25	0.73
meanwhile	1	1.00	0	–	1	0.00
since	17	0.86	26	0.86	37	0.83
(even) though	7	0.50	12	0.60	7	0.75
while	11	0.46	24	0.43	9	0.50
yet	13	0.69	8	0.69	12	0.62
Average *F1*		0.61		0.64		0.72

Table 5.13: Proportion (*P*) of labeled EN connectives as rounded percentages and F1 scores of automatic labeling (EN/DE). The total number of connectives in the three sets was 122, 165 and 176, respectively.

In addition, we tested our connective labeler on the test sets used for SMT (see Chapter 7) and report scores in Table 5.13 for each connective and globally. Given that no ground-truth labeling is available, we have manually scored the correctness of the labels for all connectives as output by the EN/DE classifier (i.e. with the respective `Translational` feature).

Table 5.13 confirms that connectives such as *since* and *yet* are rather easy to classify, while others like *while* and *however* show lower scores and varying performance. Their varying frequency in a text clearly affects the overall labeling performance: nt2008+sy2009, with the lowest average F1 score, has fewer instances of *since* and the most occurrences of *however*, while nt2010 has more occurrences of *since*, fewer of *however*, but the most of *while*. Finally, nt2012, with the best labeling performance, has the most occurrences of *since*, about the same amount of *however* as nt2010, but much fewer of the difficult *while*. Besides EN/DE, we compared the classifiers for EN/DE with those for EN/FR and EN/IT (on nt2008+sy2009) and for EN/FR (on nt2010 and nt2012). Between language pairs, the classifiers are rather stable, e.g.

4. https://github.com/idiap/DiscoConn-Classifier

in nt2008+sy2009 with EN/DE, only two connectives change with respect to EN/FR and EN/IT. These changes are due to varying baseline translations obtained for the `Translational` feature.

This chapter illustrated how the right combination of algorithms, features and sense label sets can help to reach almost human annotation performance with automatic classifiers, at least for some of the more clear-cut connective types such as *since* and *yet*. We also showed that building a specific classifier per connective type currently is the method that reaches higher performance than trying to classify all types jointly. Performance also varies widely depending on the datasets used for testing and the actual distribution of connectives to label. The next chapter will deal with automatic classification methods for verb tense, before we move on in Chapter 7 to apply these classifiers directly in SMT systems, where it can be shown that classification performance, quite intuitively, influences automatic translation quality for these discourse units.

6 Automatically disambiguating verb tense

Along the same lines as the connective classifiers described in the previous chapter, we experimented with two approaches (with different classifiers and feature combinations) for disambiguating verb tense prior to MT: one to classify EN Simple Past verbs in narrative vs. non-narrative contexts (Section 6.1) and one that directly predicts the FR tense an EN verb should be translated to (Section 6.2). Given the temporal information needed to disambiguate verb tense classes, both classifiers use similar features. The classifiers reach performances in ranges of 0.7 to 0.85 F1 score, which we regard reliable enough to annotate verbs automatically in training data for tense-aware SMT systems [1].

6.1 Disambiguating narrativity

A first automatic classification method for verb tense is trying to find whether an English verb in Simple Past tense appears in a narrative or non-narrative context, i.e. it is a binary classification task which will also be relevant for the translation of EN SP verbs, as these can be translated to up to three tenses in French, depending on their actual context. To train this classifier we make use of the manually annotated dataset that has been described in Chapter 4, Section 4.2.2.

The 435 correctly annotated instances of narrativity (257 narrative, 178 non-narrative), after resolving the disagreements as described in Chapter 4, have been used entirely for training a maximum entropy classifier with the Stanford Classifier package (Manning and Klein [2003]). Testing was performed on a smaller and earlier manually annotated sub-portion of the corpus with the same genre distribution, consisting of 118 labeled verbs: 75 instances of narrative and 43 of non-narrative uses.

From the training and test sets we extracted the following features. First, we obtained the POS tags and syntactical ancestor categories for the verbs occurring in the instances, by parsing

1. Related published papers for this chapter are (Meyer et al. [2013]) on narrativity disambiguation and (Loaiciga et al. [2014]) for automatically predicting FR verb tense.

Model	Recall	Precision	F1	κ
MAXENT	0.76	0.71	**0.72**	**0.46**
CRF	0.30	0.44	0.36	−0.44

Table 6.1: Performance of the MaxEnt classifier on labeling narrativity. Reported are overall recall, precision, their mean by the F1 score and the $kappa$ value for class agreement.

the data with Charniak and Johnson's constituent parser (Charniak and Johnson [2005]). Furthermore, a TimeML parser (Verhagen et al. [2005], Verhagen and Pustejovsky [2008]) was used for features of temporal ordering of events in the sentences. Finally, a manually compiled list of 66 temporal markers of synchrony (e.g. *simultaneously*) and asynchrony (e.g. *before*) completed the feature set. The list was mainly inspired by the temporal connectives annotated in the PDTB, and is given in Table A.1 of the Appendix.

With these features, the MaxEnt classifier performs at 0.72 F1 score (weighted mean of precision and recall). Out of the 118 test instances, the classifier correctly annotates 90 items which corresponds to an accuracy of 76.27%. As a baseline to compare against, the majority class in the test set (narrative) would account for only 64% of correctly classified instances. The detailed scores are given in Table 6.1. Moreover, also the $kappa$ value for inter-class agreement is 0.46 with the classifier and is even a bit higher than the one obtained in the first manual annotation experiment (Chapter 4).

To further test the classifier's performance, we took the data from the *first* annotation experiment (485 items, including 133 disagreements) and resolved the disagreements by looking at the tense of the FR reference translation to set the narrative vs. non-narrative labels accordingly. When trained on such data, the classifier only performs at 0.71 F1 score and at a $kappa$ of 0.43 in the test set, even though there are more training instances overall. This confirms the score range that can be expected when trying to automatically classify for narrativity.

For further comparison we built a CRF model (Lafferty et al. [2001]) in order to label narrativity in sequence of other tags, such as POS. The CRF uses as features the two preceding POS tags to label the next POS tag in a sequence of words. The same training set of 435 sentences as used above was POS-tagged using the Stanford POS tagger (Toutanova et al. [2003]), with the `left3words-distsim` model. We replaced the instances of 'VBD' (the POS tag for SP verbs) with the narrativity labels from the manual annotation. The same procedure was then applied to the 118 sentences of the test set on which CRF was evaluated.

Overall, the CRF model only labeled narrativity correctly at an F1 score of 0.36, while $kappa$ had a negative value signaling a weak inverse correlation. Therefore, the temporal and semantic features within the MaxEnt classifier are useful and account for the much higher performance of MaxEnt, which is why this model will be the one to incorporate into the SMT experiments described in Chapter 7.

Ref/Sys	narr	non-narr	Total
narr	67	8	75
non-narr	20	23	43
Total	87	31	118

Table 6.2: Confusion matrix for the labels output by the MaxEnt classifier (Sys) versus the gold standard labels (Ref).

We further study the MaxEnt classifier by providing the confusion matrix of the automatically obtained labels for the instances in the test set, in Table 6.2. It appears that labeling non-narrative uses is much more prone to errors (46.5% error rate) than narrative ones (10.7% errors). This is due to the fact that there were more instances of narrative usages in both, the training and the test data.

These classification experiments (see more details in Meyer et al. [2013]) have shown the difficulty to get correct predictions for the tense translation of EN Simple Past into French via a ±narrativity feature.

In the following section, we present a more direct approach, where a classifier attempts to predict the FR tense an EN verb should be translated to.

6.2 Automatically predicting French verb tense

An alternative approach to disambiguate for [±narrativity] is to annotate, onto the English verbs from a parallel corpus, the French tense they are translated to, and then to use this data to train and test a tense translation predictor, to be combined later with MT (Loaiciga et al. [2014]).

6.2.1 Features

For this method, as we had a 10-way classification problem (see Section 6.2.2 below), we implemented a larger number of more complex features. These are described in the following. To obtain these features, we apply a series of processors on the English texts, in the following order:

— dependency parsing (Henderson et al. [2008])
— Tarsqi toolkit for TimeML annotation (Verhagen and Pustejovsky [2008])
— Senna for syntactical parsing and semantic role labeling (Collobert et al. [2011])

All three outputs contain features that are helpful for verb tense disambiguation: from dependency parses and semantic role labeling, features such as subject, object and other constituents and clausal relations that are governed by the verb can be found and point to its tense,

and the Tarsqi toolkit has proven to provide valuable information on the temporal ordering of events in a text, as was already shown above for connectives. Overall, we extract the following features from the dependency parses, where not otherwise stated.

Verb The English word form of the verb to classify as it appears in the text.

Verb Word Forms We not only extract the verb form to label, but also all other verbs in the current sentence, and build 'bags-of-verbs' – i.e. the value of this feature is a chain of verb word forms as they appear in the text.

Position The numeric word index position of the verb in the sentence.

POS tags The POS tags for all words in the sentence are generated and output by the parser. We concatenate them to one feature value and normalize it to first five POS tags only in order to better generalize over the many possible values. As a separate attribute, we enchain the POS tags of the occurring verbs only, i.e. all POS tags such as VB, VBN, VBG etc. as they appear after the parsing. We enchain all verbs per sentence, and reduce them to the first five values as well.

Syntax Similarly to POS tags, we get the syntactical categories and tree structures for the sentences from the Senna syntactical parses and reduce them to the first five syntactical categories appearing in the tree (such as S, NP, VP, etc.).

English Tense Inferring from the POS tag of the English verb to classify, we apply a small set of rules to obtain a tense value out of the following possible attributes: The dependency parser outputs verbal tags as follows: *VB* (infinitive), *VBG* (gerund), VBD (verb in the past), VBN (past participle). Depending on the actual sequence and occurrence of these tags, one can infer the actual English verb tenses, applying the a small set of 25 rules that was described in Section 4.2.2.

Temporal Markers With the same hand-made list of temporal discourse markers as was used for the narrativity feature (see the Appendix of the thesis), we detect whether such markers are present in the sentence and use them as concatenated bag-of-word features.

Temporality of the Markers In addition to the actual marker word forms, we also indicate in our list whether a marker rather signals synchrony or asynchrony or both (such as for *meanwhile*). This additional discrete feature has thus three possible values: *s*, *a*, *a/s*, which are extracted when a marker is detected based on the values in Table A.1 in the Appendix.

Temporal Ordering The TimeML annotation language tags events and their temporal order (FUTURE, INFINITIVE, PAST, PASTPART etc.) and verbal aspect (PROGRESSIVE, PERFECTIVE etc.) and can be obtained automatically, with a precision of about 0.8 F1 score, by the Tarsqi toolkit (Verhagen and Pustejovsky [2008]).

Dependency Tags Similarly to the syntax trees of the sentences with verbs to classify, we capture the entire dependency structure via the above-mentioned dependency parser. Again, we only consider the five first dependency tags in the sentences containing the verb.

Semantic Roles The Senna parser not only allows to easily identify the head verbs of the sentences but also outputs, for each verb, its semantic roles and those of its context. As feature value, we use the one semantic role tag for the verb, which is encoded in the standard IOBES format [2] and can e.g. be of the form S-V or I-A1 (indicating the sentence (S) head verb (V) or a verb belonging to the patient (A1) in between a chunk of words (I)).

After having analyzed the above features in a MaxEnt model for predicting different sets of FR tenses (Section 6.2.2), we noted poor performance when trying to automatically predict the FR tenses of *Imparfait* and *Subjonctif*. Because these two tenses are also among the most difficult to translate with a baseline SMT system, we added two specific features to better find these two tenses. Given that these two tenses would be annotated with higher performance, the translation quality for sentences containing them, would improve as well (Section 7.8.2).

Both features were implemented after having analyzed cases in the development set, already annotated for FR verb tense.

Feature for *Imparfait* From corpus analyses with reference and baseline translations of EN verb tenses translated to FR ones, we knew that baseline systems have difficulties in finding the FR *Imparfait* tense and more often generate *Passé Composé* for EN Present Perfect and Simple Past, for example. We therefore use a small list of possible identifiers when the *Imparfait* should be the appropriate translation.

— relative pronoun + Simple Past tense in EN: we detect whether there are relative clauses starting with a pronoun such as *who, what, which, where, why* which are then followed by a verb in EN Simple Past tense
— Simple Past tense + adverb: we look further whether EN Simple Past verbs are followed by adverbs such as *repeatedly, constantly* etc. that therefore point to *imperfective* usage in FR, as the *Imparfait* merely is a tense for describing ongoing states in the past
— the third detector is used to find indirect speech with combinations of the preposition *that, as*, adverbs and the verb *said*, followed by another verb, which is then likely to be

2. The IOBES scheme indicates if each token is **I**nside a block, **O**utside, marks the **B**eginning or **E**nd of a block or if it constitutes a **S**ingle one.

translated to *Imparfait*, as in indirect speech non-narrative state descriptions seem to be more frequent (at least in the development set at hand and its newswire genres)

When the above words are found, the feature value is a binary value of yes/no that points to a likely *Imparfait* usage or not.

Feature for *Subjonctif* Similarly as with the FR *Imparfait*, a baseline SMT system is hardly capable of generating the FR *Subjonctif*. The latter is a FR mood that is used in subordinated verb constructions that express belief or unreal events. Often, the triggering main clause is too far away to be captured by a phrase-based system. We therefore added an additional feature with a binary value as well, that should point the classifier toward labeling the current verb as being likely to be translated to FR *Subjonctif*. The heuristics for finding this value were the following:

— *Subjonctif*-triggering words: In FR there are a couple of verbs and adjective expressions that, when followed by the relative initializer *que* (EN: to, that), trigger subjunctive mood: *souhaiter, espérer, supposer* etc. which express unreal or believed states. We use a small list of 15 EN verbs and adjectives that could be translated to *Subjonctif*: *so...that, (ensure, delighted, clear, vision, way, hope, good, expect, except, pleased, forward)...to, that*

When these triggers are found, the feature value then is a binary value of yes/no that should be an indicator of *Subjonctif* usage or not.

6.2.2 Results

For predicting FR tense automatically, we made use of the large gold-standard training set that is shown in Table 6.3 (and was explained in Chapter 4, Section 4.2.2 and Table 4.5) for training a maximum entropy classifier. We built three FR tense prediction models that cope with different levels of tense label granularity. The latter in turn has consequences on the SMT system and its output quality when translating labeled, i.e. predicted tenses. Classifier testing was performed on the held-out test set and features for *Imparfait* and *Subjonctif* have been found by looking through the held-out development or tuning set as listed in Table 6.3.

Sub-corpus	Number of sentences
Training	196 140
Tuning	4 000
Testing	3 000
Total	203 140

Table 6.3: Datasets for English/French verb tense prediciton.

We tested the MaxEnt models with the above-mentioned features several different sets of FR tenses as classes in order to maximize performance for the automatic translation task. Such

Configuration	F1 (cv)	F1 (test)
ALL-CLASSES	0.75	
9-CLASSES	0.85	0.83
EXTENDED	0.85	0.83

Table 6.4: Performance of the MaxEnt models on predicting FR tenses. Reported are the micro-averaged F1 scores (for 10-fold cross-validation in training (cv) and scores from testing (test) for different model configurations and data sets.

FR tense predictors should correctly output the FR tense label in as many cases as possible in order to not distort translation quality because of wrongly assigned tense labels.

ALL-CLASSES Using the full list of possible FR tenses, there are 10 classes:

présent, passé composé, imparfait, plus-que-parfait, passé simple, passé récent, passé antérieur, impératif, subjonctif, OTHER

We grouped a number of FR tenses that were often wrongly output within the semi-automatic procedure to generate the training and development data (errors were most often due to the MORFETTE tool that has difficulties on correctly annotating FR future tense). Among these OTHER tenses are the following: *Futur, Conditionnel, Futur-Conditionnel* and *Futur Proche*, which make up for 40% of the data.

9-CLASSES As the OTHER class is very frequent in number, and errors for the above-mentioned reasons were to be expected, we generated a MaxEnt model that did not include this class in order not to bias the classification results.

EXTENDED Besides the two models described above we built a third one (over the same data size as for the 9-CLASSES model) to account for the two tenses that were most difficult to annotate and for which considerable EN/FR translation divergencies exist. As was mentioned in Section 6.2.1, we extended the MaxEnt model with two specific features to better predict the *Imparfait* and *Subjonctif* tense. Our last configuration (EXTENDED) here is therefore one where we use the 9 CLASSES (all FR tenses except OTHER), but trained with two additional features.

The classification results with the three models are listed in Table 6.4. F1 scores are listed for 10-fold cross-validation on the entire training set and, when relevant (i.e. when considered for the translation task), also on the test set.

The F1 scores show that the OTHER class is indeed problematic for the overall classification performance, because it negatively influences the scores of rather infrequent tenses (such as

FR tense	9-CLASSES		EXTENDED	
	F1 (cv)	F1 (test)	F1 (cv)	F1 (test)
Imparfait	0.48	0.40	0.47	0.44
Passé Composé	0.77	0.73	0.76	0.72
Impératif	0.29	n/a	0.24	n/a
Passé Simple	0.16	n/a	0.09	n/a
Plus-que-Parfait	0.55	0.36	0.51	0.25
Présent	0.92	0.91	0.91	0.91
Subjonctif	0.33	0.16	0.29	0.17
Passé Récent	0.16	n/a	0.22	n/a

Table 6.5: Performance of the MaxEnt models on predicting specific FR tenses. Reported are the F1 scores per class for 10-fold cross-validation and on the test set. Some tenses (n/a) were not occurring in the test set.

Passé Antérieur and *Impératif*) in the training set. As soon as as this class is separated out from classifcation, performance can reach up to 0.85 F1 score and stays above 0.80 even in the test set.

For the EXTENDED model the overall performance stays the same, not revealing influence of the two new features. We therefore also performed an analysis per tense class in Table 6.5 that lists the F1 score that was obtained on each specific class. Note that FR tenses not listed were not occurring in the test set.

The EXTENDED model, based on *Imparfait* and *Subjonctif* features does not improve in cross-validation performance, but in the test set, the two tenses have slight gains of 0.04 and 0.01 F1 score, respectively. We will test these two classifier models thoroughly for their effect on tense-aware SMT systems in Chapter 7.

7 Statistical machine translation with discourse labels

In this chapter of the thesis, we present SMT methods and experiments that make use of the discourse connective and verb tense classifiers described above. As we deal with statistical MT models exclusively (as opposed to rule based ones), there is no straightforward solution on how to use linguistic information in the translation and/or language models. Although recent work has demonstrated some advantages of statistical syntactical or hierarchical translation models over phrase-based ones (as e.g. described in Section 3.3.2), in the case of lexicalized, unstructured labels as those we assign here to discourse connectives and verb tenses, the phrase-based SMT approach offers advantages in terms of robustness and simplicity of integration.

The chapter starts with a description of three oracle SMT experiments, where we do not use automatically assigned labels, but directly the gold standard ones assigned through manual annotation, either from the PDTB or from our own efforts over the Europarl corpus. These experiments provide an indication of the upper bound performance by which connective and verb tense labels can actually improve translation quality (Sections 7.1.1 and 7.1.2) [1].

The experimental settings (in the entire chapter) always compare an MT system that was trained on discourse information, based on the available labels, with a baseline one, which was built over the same amount of data, but learning from plain text only, not incorporating any discourse features. The experiments are therefore not comparable over the entire chapter, but for each of them a baseline counterpart to the modified systems is always available for comparison.

When evaluating translations automatically by computing BLEU scores, we show that the scores remain stable across the modified and the baseline SMT systems, due to the very few changes that are performed per sentence (connectives and verb phrases are usually no longer than three words). We therefore resort to manual evaluation for most experiments described

1. The oracle experiments were collaborations with Lucie Poláková of Charles University, Prague (for Czech translation evaluation) and with Sharid Loáiciga, intern at Idiap and PhD student at the University of Geneva (for the system based on oracle verb tense labels).

in this chapter and we count the number of connectives or verb phrases that were improved, how many of them stayed the same and how many were degraded.

In Section 7.1, using oracle settings and manual evaluation, we show that the translation of discourse connectives improves in ranges of 17%-21%, while the BLEU scores remain generally stable. As there are far more verbs than connectives in sentences, a significant gain in BLEU is observed (+0.5 points) with oracle labels on verb phrases. The translation of tenses improves by about 25%, when assessed by humans, while lexical choice and person-number agreements remain at the level of a baseline SMT system.

Sections 7.2 to 7.5 illustrate several ways to use automatically assigned labels for SMT. First, following a simple rule-based idea, we search within a baseline translation model (i.e. a phrase table) and based on a dictionary for EN connectives, for connectives that contain a valid explicit FR connective in the phrase pair to which a disambiguating sense label is assigned. Additionally, we augment the translation probability score for these pairs. The procedure improves connective translation by 14%, but it must be repeated from scratch for each new target language.

Following these experiments, we then make direct use of the connective classifiers that allow to assign the labels *prior* to training the systems, either onto the EN connective word forms, or with their probabilities that point to the most likely labels. Baseline and modified SMT systems can also be combined by letting the label-aware system translate all classifiers for which the label was assigned with high confidence, and submitting the others to a baseline system. Depending on the settings, improvements in ranges of 4% to 18% are achieved.

We then argue that simply post-editing (correcting) erroneously translated connectives in the SMT output (Section 7.6), does not yield the same improvements as the factored translation models, which we present in Section 7.7, and which are our most principled proposal for discourse-aware SMT, reaching improvements in translation quality of 2 to 8.5 points in terms of the ACT reference-based automatic metric for connectives [2].

Having learned from the connective experiments, we similarly built factored models to better translate verb tense, based on automatically assigned narrativity information on EN Simple Past verbs (Section 7.8.1) or automatically assigned FR tenses on all EN verbs (Section 7.8.2). As verbs are far more frequent than connectives, the translation improvements here are also measurable with the BLEU score, which increases by about 0.2 points. Still, we performed manual evaluation of translations in order to quantify the specific improvement on verb tenses, and found that tense conjugation improved in a range of 10% to 20%, lexical choice improved by up to 3% and the overall correctness of verb phrase translations augmented by up to 9%.

2. SMT with automatically annotated connectives largely profited from collaboration within the COMTIS project. We published the first experiments early in the PhD work (Meyer [2011] and Meyer and Popescu-Belis [2012]). Then, Andrea Gesmundo (a COMTIS PhD student in Geneva at the time) contributed to the hierarchical model in Section 7.7.1, published in (Meyer et al. [2012]). Najeh Hajlaoui (a COMTIS postdoc at Idiap at the time) built the models with Arabic as a target language (Section 7.7.2) and the post-editing method (Section 7.6), both submitted to a journal (Meyer et al. [2014]).

7.1 Oracle experiments

In order to test how much translation quality would actually improve by using the label information from the connective and tense disambiguation modules, we performed a series of experiments in which we directly use manually annotated and therefore gold-standard labels in the training and testing stages for discourse-aware SMT systems. In other words, the goal was to assess the translation improvement if perfect labels would be available. In the following subsections we describe these oracle experiments, for connectives (Section 7.1.1) and verb tenses (Section 7.1.2).

7.1.1 SMT with oracle disambiguation of connectives

With discourse connectives, we had two possibilities in order to test for oracle SMT performance. On the one hand, we had the manual annotation of about 2000 sentences in the Europarl corpus for English/French with up to seven types of connectives (Chapter 4). This is nowhere near the amount of sentences that is normally needed to train an SMT system (corpora with hundreds of thousands of sentences are usually used). Still, as system building with manual annotation is easy, it was worth trying whether changes in translation quality would occur with such few annotated connectives.

On the other hand, as was mentioned earlier, the PDTB provides annotation for 18'459 explicit discourse connectives in English and of up to 100 connectives types. However, the only human translation of the corpus that is available is into Czech, as the Prague Czech-English Dependency Treebank (PCEDT) (Hajič et al. [2011]) [3]. This provides a human translation of the entire Wall Street Journal Corpus (sections 00-24, approximately 50'000 sentences and 1'000'000 tokens). This dataset has the advantage that the entire PDTB annotation can directly be used for training and testing an EN/CZ SMT system.

SMT with manual annotation of connectives in Europarl

In order to test SMT systems that directly make use of manually annotated discourse connectives, we took our Europarl datasets of the five EN connectives *although, even though, since, though, while* (Table 4.4) and directly integrated them with the rest of the Europarl corpus. The discourse relation labels were directly concatenated onto the EN connective word forms. This combination method will also be used, along with others, for automatically assigned labels and is further described in Section 7.3. This resulted in the following overall data for the SMT training procedure with the Moses SMT toolkit (Koehn et al. [2007]): Europarl v5, EN/FR (only direct translations, see 4.1.2), 346,803 sentences, minus all 8,901 sentences containing one of the 5 connective types, plus 1,147 sentences with manually sense-labeled connectives. All data was tokenized and lowercased using the Moses tools. For MERT tuning (Och [2003])

3. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T08

the development set was News Commentary 2009 [4], 2,051 sentences, minus all 123 sentences containing one of the 5 connective types, plus 102 sentences with manually sense-labeled connectives. Testing was performed on 35 sentences from News Commentary 2007, with 7 occurrences for each of the 5 connective types, manually labeled. The 5-gram language model was built over the entire FR side of Europarl v5, using SRILM (Stolcke et al. [2011]).

Over this test set, as was expected, the BLEU scores of our modified system are close to those of the baseline unmodified system (actually slightly lower), due to the very few changes to connectives only: 41.58 vs. 42.77 points, also confirmed when the BLEU scores are bootstrapped (as explained in Section 4.3): 42.38 vs. 43.54. However, when manually evaluating the 35 connective translations, a clear improvement can be seen: our modified system translated the 5 connective types better in 32% of the cases, similarly in 57% of the cases, and only 11% are degraded. This oracle experiment has been described in (Meyer and Popescu-Belis [2012]).

SMT with the PDTB annotation of connectives

For translating into Czech the English connectives that have been annotated over the PDTB, we built two complete SMT systems, by using the label concatenation method (see Section 7.3) and two different granularity levels for discourse relations in the PDTB hierarchy. For SYSTEM1, we inserted, on the English texts that are translated in the PCEDT data, the full sense labels from the PDTB, which have up to three sense levels and allow for composite labels indicating that two senses hold at the same time. SYSTEM1 therefore operates on a total of 63 observed sense labels for all discourse connectives. For SYSTEM2, we reduced the labels to those from the first and second levels of the PDTB sense hierarchy only, and simplified the composite labels by discarding all but the first sense for the instances annotated with multiple labels (though they are not necessarily less important). This reduced the set of senses for SYSTEM2 to 22.

The procedure is exemplified below. The first EN sentence (from WSJ section 2300) contains a complex PDTB sense tag that is used for SYSTEM1. For SYSTEM2, we have reduced the sense of *when* to <CONTINGENCYCONDITIONGENERAL>. Sentence 2 (from WSJ section 2341) contains two already simplified sense tags. The original PDTB sense tags for *meanwhile* and *as* were respectively <COMPARISONCONTRASTJUXTAPOSITION> and <CONTINGENCYPRAGMATICCAUSE-JUSTIFICATION>, where JUXTAPOSITION and JUSTIFICATION were dropped because they are from the third level of the PDTB sense hierarchy.

1. Selling snowballed because of waves of automatic "stop-loss" orders, which are triggered by computer **when**<CONTINGENCYCONDITIONGENERAL-TEMPORALASYNCHRONOUSSUCCESSION> prices fall to certain levels.
2. Meanwhile<COMPARISONCONTRAST>, analysts said Pfizer's recent string of lackluster quarterly performances continued, **as**<CONTINGENCYPRAGMATICCAUSE> earnings in the quarter were expected to decline by about 5%.

4. Distributed by the Workshop on Statistical Machine Translation at http://www.statmt.org/wmt12/.

In order to train SMT systems up to a reasonable quality level, we still need to combine the PCEDT texts (50k sentences) with other parallel resources such as the EN/CZ parts of the Europarl corpus. This results in a mixture of labeled (PDTB) and unlabeled (Europarl) discourse connectives in the English data. The Czech PCEDT translation of the PDTB does not contain any labels. We additionally checked system performance on the PDTB test set (section 23) with labeled discourse connectives only (see Table 7.1) for which the unlabeled ones in the model do not pose a problem, as the SMT decoder can only search the phrase table for phrase pairs with labeled connectives (because only labeled ones are present in the test set to translate). The data used to build three SMT systems with the Moses SMT toolkit (Koehn et al. [2007]) was divided as follows.

The BASELINE system exactly has the same amount of sentences, but no sense labels. All data was tokenized and truecased by using the Moses tools. The language model, the same for BASELINE, SYSTEM1 and SYSTEM2, was built using SRILM (Stolcke et al. [2011]) with 5-grams over Europarl and the news data sets 2007-2011 in CZ, as distributed by the Workshop on Machine Translation [5]. The systems were tuned by MERT (Och [2003]) as implemented in Moses.

Training data: Europarl v7 (645,155 sentences) + PDTB sections 02-21 (41,532 sentences; 15,402 connectives)

Tuning data: newstest 2011 (3,003 sentences) + PDTB sections 00, 01, 22, and 24 (5,260 sentences; 2,134 connectives)

Testing data: (1) newstest 2012 (3,001 sentences) + PDTB section 23 (2,416 sentences; 923 connectives); and (2) PDTB section 23 only (2,416 sentences; 923 connectives). This division of PDTB into training, development and test data is the same as the one recommended in the PDTB annotation manual and used in Chapter 5 for automatic classification experiments.

Table 7.1 provides the BLEU scores for the BASELINE and SYSTEMS 1 and 2 on the two test sets. For discourse connectives, global reference-based evaluation metrics such as BLEU do not reveal much of a system's performance, as often only one or two words, i.e. mainly the discourse connective itself, are changed. When a candidate translation however contains a more accurate and correct connective than the baseline's output, the candidate's output is often more coherent and readable. Still, in order to obtain reliable automatic evaluation scores, we executed five runs of MERT optimization for each configuration, and averaged the scores using the MultEval tool as stated in Section 4.3. In terms of such averaged BLEU, both SYSTEM1 and SYSTEM2 perform at the same BLEU scores.

In order to show that the labeling of discourse connective still can affect the BLEU score, we randomized all connective sense tags in the PDTB test section 23 and translated again five times (with the weights from each tuning run) with both SYSTEM1 and SYSTEM2. With

5. http://www.statmt.org/wmt12/

randomized labels, both systems perform significantly worse ($p = 0.01$, marked with a star in Table 7.1) than the BASELINE, with an average performance loss of 0.6 BLEU points. Moreover, some sense tags might still have been correct by chance. This is a strong indication that having correct discourse sense tags is important and of direct influence on translation performance.

Test set	System	BLEU
newstest 2012 + PDTB section 23	BASELINE	17.6
	SYSTEM1	17.6
	SYSTEM2	17.6
PDTB section 23 only, with random labels	SYSTEM1	20.8*
	SYSTEM2	20.8*
PDTB section 23 only, with gold labels	BASELINE	21.4
	SYSTEM1	21.4
	SYSTEM2	21.4

Table 7.1: BLEU scores when testing on the combined test set (newstest 2012 + PDTB 23); when randomizing the sense tags (PDTB 23 random) and on PDTB section 23 only; for the BASELINE system and the two systems using PDTB connective labels; SYSTEM1: complex labels, SYSTEM2: simplified labels. When testing on randomized sense labels (PDTB 23 random), the BLEU scores are significantly lower than the ones on the correctly labeled test set (PDTB 23), which is indicated by starred values.

We now turn to the human analysis of the translation output by SYSTEM2, which we selected over SYSTEM1 because it reached the highest scores observed in some of the tuning runs before averaging. Two linguists, which are native speakers of Czech, went through three random samples of SYSTEM2 translations from WSJ section 23, namely sentences 1–300, 1000–2416 and 1024–1138. In these sentences, there were 680 observed connectives. The judges counted the translations that were better, similar or worse in terms of discourse connectives in the output from SYSTEM2 compared to the BASELINE system. The consolidated counts over the three samples are given as $\Delta(\%)$ in Table 7.2. A translation was counted as being correct when it generated a valid CZ connective for the meaning of the source EN connective, without grading the rest of the sentence.

Overall, it was found that the number of translations improved by SYSTEM2 in comparison to the BASELINE is in the same range as those that were degraded, though clearly smaller than the number of discourse connectives that were translated correctly by both the BASELINE and SYSTEM2 (the vast majority). In very few cases, both systems translated the discourse connectives incorrectly (respectively 7% and 2% for the data sets in Table 7.2).

SYSTEM2 appeared to systematically repeat one mistake, namely translating the very frequent connective *but* preferably with *jenže*, which is correct but rare in CZ (the primary and default equivalent for *but* in CZ is *ale*). The phrase pair *but–jenže* has received a higher weight in the translation model due to its frequency in the SMT training data, which did not have the same style than the testing data. If one disregards these occurrences, SYSTEM2 translates between 8

Configuration	Δ(%) **vs. BASELINE**			**Total (%)**
	Improved	Equal	Degraded	
sentences 1–300 and 1000–2416 630 labeled discourse connectives				
SYSTEM2	7.9	75.2	9.4	92.5
not counting 25 x *but–jenže*	8.2	80.3	4.0	92.5
				100
sentences 1024–1138 50 labeled discourse connectives				
SYSTEM2	16	76	6	98
not counting 2 x *but–jenže*	19	77	2	98
				100

Table 7.2: Performance of SYSTEM2 (simplified PDTB tags) when manually counting for improved, comparable or degraded translations compared to the BASELINE, in samples from the PDTB section 23 test set.

and 20% of all connectives better than the BASELINE.

Especially for SYSTEM1, but to some extent also for SYSTEM2, rare sense tags such as CONTINGENCYPRAGMATICCONTRAST are not often seen in the SMT training data (only 4 occurrences in the entire PDTB) and are therefore not learned appropriately. In relation to that, simply concatenating the sense tags onto the connective word forms leads to data scarcity, whereas other ways to include linguistic labels in SMT, such as factored translation models, rather account for labels as additional translation features (see Section 7.7).

The results of the oracle experiment at this point seem therefore to depend on the exact test set and the discourse connectives occurring in it. This observation is confirmed when testing the translation of automatically labeled discourse connectives in the Europarl corpus, with factored translation models, on several testsets and several target languages: the more correctly certain discourse connectives are labeled, and the higher their frequency in the test set, the better the resulting translation performance (Section 7.7.2).

7.1.2 Oracle SMT with verb tense

In order to test how much the labeling of verb tenses can improve SMT, we took the whole semi-automatically generated annotation of 203'140 sentences from the EN/FR Europarl corpus v7 (see Chapter 4) and subtracted the last 7000 sentences for tuning (4000 sentences) and for testing (3000 sentences). Note that the oracle labels on EN source verb phrases, indicating the expected tense translation into FR, were found through the automatic alignment procedure described in Chapter 4, which is not entirely devoid of errors.

Using the training and tuning sets, we built a factored translation model (Koehn and Hoang [2007]) as implemented in the Moses SMT toolkit (Koehn et al. [2007]) that learns the added tense labels as additional translation features (Section 7.7). The language model was a 3-gram one built by using the IRSTLM toolkit (Federico et al. [2008]) over Europarl v7 FR plus the FR side of the News Commentary corpus (years 2007-2011), as distributed by the Workshops on Statistical MT.

The translation system was tested on the gold labels of the test set, and later (see Section 7.8.2) on the same test set that had been labeled automatically by the classifier described in Section 6.2. The accompanying baseline system was again built over the same amount of data, not considering any labels. All data was tokenized and lowercased using the Moses tools.

When testing on gold labels, the overall translation quality improves by 0.5 BLEU points, from 27.73 (baseline) to 28.23 (factored) [6]. When examining for each FR tense occurring in the 3000-sentence test set the BLEU scores over the corresponding sentences per tense class, shown in Table 7.3, it becomes visible the that the overall improvement in translation quality is actually due to the labeled tenses, because the largest gains in BLEU scores are due to sentences that contain the EN/FR tense divergencies for past tenses (explained in Chapter 2).

FR tense	**Baseline**	**Tense-aware**	**Δ**	**Number of sentences**
Imparfait	24.10	25.32	**1.22**	122
Passé composé	29.80	30.82	**1.02**	636
Impératif	19.08	19.72	0.64	4
Passé simple	13.34	16.15	**2.81**	6
Plus-que-parfait	21.27	23.44	**2.17**	17
Présent	27.55	27.97	0.42	2618
Subjonctif	26.81	27.72	0.91	78
Passé récent	24.54	30.50	**5.96**	3
Average/Total	23.31	25.21	**1.89**	3484

Table 7.3: Comparison of BLEU scores of a baseline SMT system and an oracle, tense-aware one using gold-standard tense labels for FR verbs, assigned to EN verbs prior to translation.

In order to confirm that these improvements, as measured by BLEU, are due to the verb phrases that have been annotated with their gold tense labels, three annotators examined a sample of 652 verb phrases in 313 sentences, as output by the oracle SMT system in the test set. The evaluation criteria for each verb phrase, for the oracle, tense-aware SMT system and its baseline counterpart, were the following:

— Tense/Mode/Aspect (TAM): is the TAM of a verb phrase correctly translated by a system, and if correct, is it the same as in the reference translation?
— Lexical choice: is the lexical form of a verb phrase translated by a system correct or wrong, and if correct, is it the same as in the reference translation?

6. The BLEU scores for this experiment were computed by the script multi-bleu.perl provided with Moses.

— Agreement (Person/Number): Is the subject-verb agreement in terms of person and number correct for the verbs output by the system?

The results, calculated as absolute counts and percentages, are given in Table 7.4.

	TAM			Lexical choice			Agreement ok		Total VPs
	Wrong	Right		Wrong	Right		Yes	No	
		≠ ref	= ref		≠ ref	= ref			
Baseline	206	61	387	47	267	340	536	118	654
	32%	9%	59%	7%	41%	51%	82%	18%	100%
Oracle	52	39	563	60	247	347	532	122	654
	8%	6%	86%	9%	38%	53%	81%	19%	100%

Table 7.4: Manual evaluation of a baseline and an oracle tense-aware SMT system using gold-standard tense labels for FR verbs, assigned to EN verbs prior to translation, against reference translations.

In terms of tense translation, the oracle system outperformed the baseline with respect to TAM features (+27% of better translations, as seen from the percentages in 7.4). The lexical choice and the agreement counts, on the other hand, did not change much between these configurations. We further analyzed the manual evaluation results per translated tense which confirmed that infrequent target tenses are better generated by the tense-aware system, for instance the Passé Simple (+66.6% with respect to the baseline) and the Passé Récent (+100% with respect the baseline). These tenses, however, are of rather low frequency. By contrast, the FR Imparfait and the Subjonctif tenses, which are of higher frequencies (25% and 12% respectively), also reveal that English tenses with a real translation ambiguity were better translated by the tense-aware system. For instance, most of the Present Perfect EN VPs were translated as Passé Composé by the baseline system, since this is the most frequent translation with about 60% of the translations of Present Perfect EN VPs (see Table 2.1). The tense-aware model correctly boosted the number of translations into the FR Imparfait tense (+47% with respect to the baseline). Similarly, for the FR Subjonctif, the improvement amounted to +56% with respect to the baseline system.

Starting with the next section, we discuss, in contrast to the oracle experiments described so far, fully automated methods that directly use the output of the discourse connective and verb tense classifiers, in order to make available many more labeled instances for training and testing SMT systems. We experimented with six different methods to achieve this goal, and we will describe each of them in a separate subsection below.

7.2 Phrase table modification

A first method to make use of discursive labels for SMT is to search for occurrences of English connectives in the phrase table that is generated during the training stage of a phrase-based SMT system. When, in a phrase pair, the target language connective indicates only one of the possible senses of the English connective, then the sense label is added to the English

connective, and the lexical probability feature score of the pair is increased. This means that neither a label from manual nor from automatic annotation as described above is used, but is created by searching through and modifying the phrase table based on a small dictionary of possible connective translations and their senses, as known from the target connectives. Overall, this method amounts to pruning wrong translations from the phrase table and leads to small improvements in translation at the cost of rule-based phrase table editing. Nevertheless, given such dictionaries, the method is cheap to implement and to test.

For the EN/FR language pair, for example, for every phrase table entry in which *while* is translated with an FR connective that clearly expresses temporality (e.g. *pendant que, tout en, ...*, *while* is changed into *while*_TEMPORAL. Or, for the entries in which *while* is translated as e.g. *bien que, même si, ...*, the lexical entry is changed into *while*_CONCESSION. We increased the lexical probability scores to the maximum value (i.e. 1) for such modified phrases. However, when the target entry does not correspond to a unique sense (i.e it does not solve the ambiguity of the source entry), no modification is made. This means that during decoding (testing) with labeled sentences, these entries will never be used. The following example gives an idea of the changes in the phrase table of an EN/FR Moses SMT system:

Original:
and the commission , while preserving ||| et la commission tout en défendant ||| 1 3.8131e-06 1 5.56907e-06 2.718 ||| ||| 1 1
and while many ||| et bien que de nombreuses ||| 1 0.00140575 0.5 0.000103573 2.718 ||| ||| 1 1

modified:
and the commission , *while*_TEMPORAL preserving ||| et la commission tout en défendant ||| 1 **1** 1 **1** 2.718 ||| ||| 1 1
and *while*_CONCESSION many ||| et bien que de nombreuses ||| 1 **1** 0.5 **1** 2.718 ||| ||| 1 1

For building such a modified system, we trained a baseline phrase-based model using the Moses SMT toolkit (Koehn et al. [2007]) and then modified its phrase table after the training stage, while keeping an unmodified copy for the baseline system.

The data for training and tuning was the same as in the oracle experiment described above (Section 7.1.1), namely Europarl v5 EN/FR with 346,803 sentences for training, and NC 2009 with 2,051 sentences for tuning, and the same 5-gram language model obtained from Europarl v5 for FR. All data was tokenized and lowercased using the Moses tools. After producing the phrase table, we introduced the sense tags from a simple dictionary for the following 5 EN connective types: *although, even though, since, though, while.* One of the two test sets was the same as above, i.e. 35 sentences with 7 occurrences for each of the 5 connectives types, which were hand-labeled in order to be considered by the modified phrase table at decoding time. A second and much larger test set, with 10,311 sentences containing one of the 5 connectives types, was automatically labeled with the classifier described in Section 5.6 above. The results of the two SMT systems are shown in Table 7.5.

In the first test set, the translations of 29% of the connectives are improved by the modified

MT system	Conn. in MT test data			Δ *Conn.* (%)			BLEU scores	
	Occ.	Types	Labeling	+	=	–	Standard	Bootstrap
Modified phrase table	35	5	manual	29	51	20	39.92	40.54
Baseline	35	5	–	–	–	–	42.77	43.54
Modified phrase table	10,311	5	auto	34	46	20	22.13	23.63
Baseline	10,311	5	–	–	–	–	22.76	24.34

Table 7.5: Performance of a baseline MT system and of a system using a modified phrase table, on two datasets: one labeled manually and one automatically, in terms of BLEU scores (including bootstrapped ones) and variation in the translation of individual connectives (Δ *Connectives*, as a percentage).

system, while 20% are degraded and 51% remain unchanged – thus reflecting an overall 10% improvement in the translations of connectives (Δ *Connectives*). However, for this test set, the BLEU score is about 3 points below the baseline SMT system.

For the second test set, the BLEU score of the modified system is in the same range as the score of the baseline one. As for $\Delta Connectives$, because it was not possible to score manually all the 10,311 connectives, we sampled 35 sentences and found that 34% of the connectives are improved, 20% are degraded and 46% remain unchanged, again reflecting an improvement in the translation of connectives. This shows that piping automatic labeling and SMT with a modified phrase table does not degrade the overall BLEU score, while increasing $\Delta Connectives$. The experiments have been described in (Meyer and Popescu-Belis [2012]).

Phrase table modification has the big disadvantage that it cannot handle the case of ambiguous connectives in the target language, or of phrases for which the sense on the source side cannot be identified. Moreover, the rule-based editing of the phrase table must be repeated for each language pair and connective type. When comparing to the experiment that assigns oracle labels to the connectives in all SMT training/tuning/test data (Section 7.1.1), one can also see over all scores, for Δ *Connectives* and BLEU, that phrase table modification reaches lower scores, due to the imprecise assignment of labels when an unambiguous explicit target connective is present in the current phrase, whereas oracle labels improve translation quality to a larger extent, even when only a small amount of such labels are present.

Moreover, modifying entries in a translation table does not take advantage of the statistical methods underlying SMT training and decoding, which is why we move on to methods that actually learn the labels as translation features during the training stage of SMT systems.

7.3 Concatenating labels to word forms

A method that was used in a number of studies, including ours [e.g. Chan et al., 2007, Meyer and Popescu-Belis, 2012, Meyer and Poláková, 2013], appends linguistic labels directly to word forms, in our case connectives, thus creating new word forms that are learned in a translation

English Sentence	French Sentence
Baseline	
since this network hadn't changed in the course of human evolution, they sa , it might well be a cause of cross-species sensations.	**comme** ce réseau ne s'est pas beaucoup modifié au cours de l'évolution humaine, il pourrait tout à fait participé de sensations transversales en tout genre.
while nearly every cell phone can play mp3 files, no mp3 player can make phone calls.	**alors que** pratiquement tous les mobiles savent lire des fichiers mp3, aucun lecteur mp3 ne sait téléphoner.
Label concatenation	
since-cause this network hadn't changed in the course of human evolution, they say, it might well be a cause of cross-species sensations.	**comme** ce réseau ne s' est pas beaucoup modifié au cours de l'évolution humaine, il pourrait tout à fait participé de sensations transversales en tout genre.
while-contrast nearly every cell phone can play mp3 files, no mp3 player can make phone calls.	**alors que** pratiquement tous les mobiles savent lire des fichiers mp3, aucun lecteur mp3 ne sait téléphoner .

Figure 7.1: Examples of parallel sentences in EN/FR as training input to a baseline (above) and discourse-aware SMT system (below) using sense labels concatenated to connectives.

model and can be used when translating. For training, gold-standard labels (fewer but more accurate) or automatically-assigned ones can be used. Still, the tuning and test data need to be labeled in order to trigger the search for labeled connectives in the translation model during decoding.

The method automatically labels the connectives on the source side of the SMT training data, as in Chapter 5. The labels are thus present before constructing the phrase table, which is generated as in the baseline SMT system, with the difference that the labels are directly learned with the corresponding connectives present in the source and target language texts. Figure 7.1 shows an example of the input to a baseline and, respectively, a discourse-aware SMT system.

We experimented with mixtures of manually and automatically labeled data, or only automatically labeled data. Unlike the oracle experiments in Section 7.1.1, no manual annotation is used in the testing data. Using only automatically assigned labels in training data provides a larger (but also noisier) amount of data. Still, manual annotations are not present at all in this case, except for the initial training of the classifiers (see Chapter 5). We carried out five experiments with various amounts of labeled connectives and overall training/tuning/test data in order to compare SMT system performances. The phrase-based translation models were again built by using the Moses SMT toolkit (Koehn et al. [2007]) and the same language model as above, i.e. a 5-gram language model over the entire FR side of Europarl v5, built with SRILM (Stolcke et al. [2011]). Pre-processing of the texts again involved tokenization and

SMT training	N.	Conn. in MT test data			$\Delta Conn.$ (%)			BLEU scores	
		Occ.	Types	Labeling	+	=	–	Standard	Bootstrap
Man. annotations	1	10,311	5	Cl. EU	26	66	8	22.43	24.00
Automatic	2	62	13	Cl. PT	16	60	24	14.88	15.96
Classifier PT	3	10,311	5	Cl. EU	16	66	18	19.78	21.17
Automatic	4	62	13	Cl. PT+	11	70	19	15.67	16.73
Classifier PT+	5	10,311	5	Cl. EU	18	68	14	20.14	21.55

Table 7.6: MT systems dealing with manually and automatically (PT, PT+, EU) sense-labeled connectives: BLEU scores (including bootstrapped ones) and variation in the translation of individual connectives ($\Delta Connectives$, as a percentage). The baseline scores are mentioned in the text.

lowercasing with the Moses tools. The following data sets were used for the five experiments:

Experiment 1. **Training:** Europarl v5 EN/FR (346,803 sentences), minus all 8,901 sentences containing one of the 5 connective types (*although, even though, since, though, while*), plus 1,147 sentences with manually sense-labeled connectives. **Tuning:** NC 2009 (2,051 sentences), minus all 123 sentences containing one of 5 connective types, plus 102 sentences with manually sense-labeled connectives. **Testing:** 10,311 sentences from the EN/FR UN corpus, all occurrences of the five connective types, automatically labeled with classifier 'EU' (Table 5.7).

Experiment 2. **Training:** Europarl v5 EN/FR – years 199x (58,673 sentences), all occurrences of the 13 PDTB subset connective types have been labeled by classifier 'PT' (cf. Table 5.2) (6,961 occurrences). **Tuning:** NC 2009 (2,051 sentences), all occurrences of the 13 PDTB subset connective types have been labeled by classifiers (340 occurrences). **Testing:** 62 sentences from NC 2007 and 2006 with occurrences for the 13 PDTB connective types, automatically labeled with the same classifiers.

Experiment 3. **Training, tuning:** Same as experiment 2. **Testing:** Same as experiment 1.

Experiment 4. **Training, tuning, testing:** Same as experiment 2, but all labeling done by classifier 'PT+' (see Table 5.2).

Experiment 5. **Training, tuning:** Same as experiment 4. **Testing:** Same as experiment 1.

The results of these five SMT systems on the different test sets and under the various labeling conditions are shown in Table 7.6 and analyzed in the following.

Experiment 1 is similar to the oracle condition described above (Section 7.1.1), except that the 10,311 occurrences of the five connective types in the test set were labeled automatically. In a sample of 35 sentences of the test set, 26% of all connectives were improved, 66% remained the same, and only 8% were degraded. Overall, the BLEU scores of our modified systems are similar to the baseline ones (22.43 vs. 22.76), which is also confirmed by the bootstrapped scores. Another comparison shows that the system trained on manual annotations also

outperforms the system using a modified phrase table (Section 7.2) in terms of BLEU scores (22.43 vs. 22.13) and bootstrapped ones (24.00 vs. 23.63).

For experiments 2 and 3, the BLEU scores as well as the manual counts of improved connectives are lower than in the preceding experiments because, overall, fewer training/tuning data was used – about 15% of Europarl – though overall they contain a larger amount of labeled connectives due to automatic labeling. The baseline system was built over the same amount of data, with no labels. In experiment 2, testing was performed over a slightly larger test set with 62 sentences and 13 connective types. The occurrences were tagged with Classifier PT prior to translation. Compared to the baseline system, the translations of 16% of the connectives were improved, while 60% remained the same and 24% were degraded. In experiment 3, the 10,311 occurrences of five connective types in the UN corpus were first tagged with Classifier EU. Evaluated on a sample of 62 sentences, 16% of the connectives were improved, while 66% remained the same and 18% were degraded. Despite fewer training data, in terms of BLEU, the difference to the respective baseline system is similar in both experimental settings: 19.78 vs. 20.11 for experiment 3 with automated annotation, compared to 22.43 vs. 22.76 for experiment 1 with manual annotation (these numbers are not shown in Table 7.6).

Finally, we carried out two experiments (4 and 5) with Classifier PT+, which uses as additional features the translation candidates and has a higher accuracy than PT, as shown in Section 5.4. As a result, the translation of connectives (Δ *Connectives*) is indeed improved compared (respectively) to experiments 2 and 3, as it appears from lines 4–5 of Table 7.6. Also, the BLEU scores of the corresponding SMT systems increase in experiment 4 with respect to 2, and in experiment 5 with respect to 3, and are now equal to the baseline ones (for experiment 5: 20.14 vs. 20.11, or, for bootstrapped scores, 21.55 vs. 21.55).

The results of experiments 4/5 vs. 2/3 indicate that improved classifiers for connectives also improve SMT output, as measured by Δ *Connectives*, with BLEU remaining fairly constant. This is the reason why the classifiers for further experiments will keep the translational feature, along with other new ones, to maximize the classification accuracy (as in Section 7.7). The experiments presented here were the first ones to show that the accuracy of a connective labeler has a direct influence on translation quality. We will further confirm these results by experiments in the following sections and will later make the same point for a verb tense predictor: the more accurate its predictions for tenses, the higher the translation quality for verbs (see Section 7.8.2).

When comparing manual annotations (experiment 1 and Section 7.1.1) to automated ones (as in experiments 2–5) regarding their metrics for SMT, the differences in terms of BLEU and Δ *Connectives* scores highlight an important trade-off: manually annotated data used for training leads to better scores, but noisier and larger training sets that are annotated automatically are an acceptable solution when manual annotations are not available.

Despite the improvements in connective translation quality, label concatenation introduces sparsity in the training data: for example, a connective *since* with the same phrasal context

Connective	Features	Ground truth	Predicted	Confidence
although	IN provides VBZ not_found claim VBP not_found progress NN optimistic JJ 'A,CA' ...	concession	concession	0.999
although	IN does AUX do_third do AUX do_inf yesterday NN votes NNS 'A,CA' ...	contrast	concession	0.799
although	IN has AUX have_third was AUX be_past materials NNS spillage NN 'ACA' ...	contrast	contrast	0.957

Figure 7.2: Three examples of labels assigned by a MaxEnt model for discourse relations, with feature excerpts, gold and predicted answers, and confidence scores. The second example illustrates a wrong decision, accompanied by a low confidence score.

(e.g. *since this network*) can appear once with a TEMPORAL label and once with a CAUSAL label, depending on the wider context. Hence, two phrase pairs are generated for the phrase table, whereas a baseline system would create only one. This can reduce the amount of training data for connectives, which is why we will explore other methods of integrating discursive labels in the following subsections. Still, it must be noted that creating two different phrase pairs helps disambiguating the connective usage and finding the correct target phrase, provided that connectives in the test set are correctly labeled.

7.4 System combination based on labeling confidence

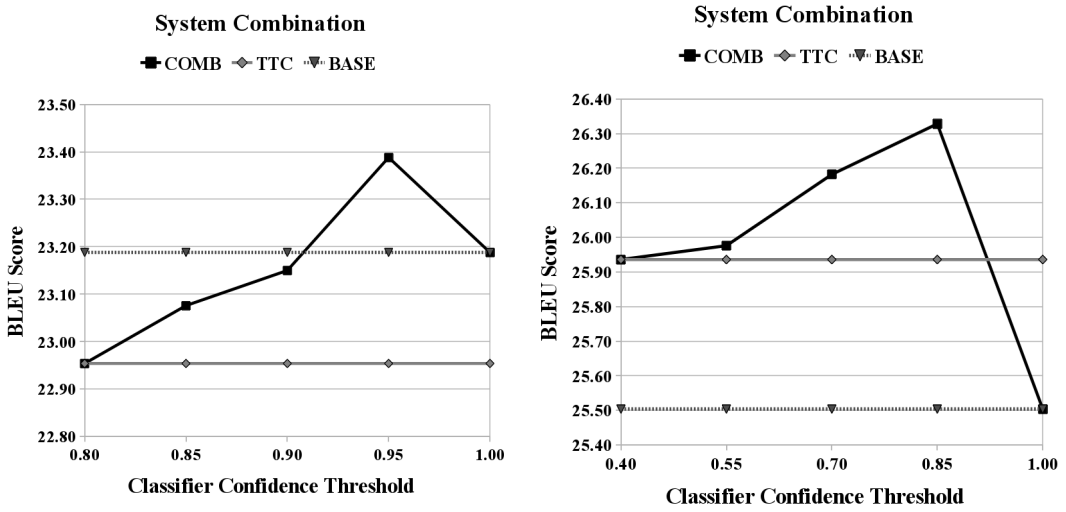
To address the situation in which an automatic classifier assigns a wrong label to a connective, which is then wrongly translated, we carried out an experiment to test the hypothesis that an SMT system dealing with labeled connectives would best be used only when the confidence of the classifier is above a certain threshold, while a generic baseline SMT system would be used for confidence values below the threshold.

The maximum entropy models described in Sections 5.4–5.6 output probabilities or confidence scores when deciding on a discourse label to be assigned to a connective, based on the corresponding features. Figure 7.2 shows an example with classifier decisions and their confidence scores. We experimented with the confidence scores of the classifier EU mentioned in Table 7.6, which assigns a confidence score between 0 and 1 to each of its decisions on the connectives' labels.

We defined a threshold-based procedure to combine SMT systems: if the confidence for a sense label is above a certain threshold, then the sentence is translated by an SMT system trained on labeled data from experiment 4 (or "tagged corpus", hence noted TTC), and if it is below the threshold, it is sent to a baseline system (noted BASE). Resulting is a combined

system, when evaluating for the joint performance of TTC and BASE (COMB).

Firstly, we considered all the 1,572 sentences from the UN test set (Experiment 1 in Section 7.3) which contained the connective *although*, labeled either as CONTRAST or CONCESSION. In Figure 7.3(a), we represent BLEU scores of the COMB system for several thresholds within the interval of observed confidence scores (0.8 to 1), along with the scores of BASE and TTC. The results show that the scores of COMB increase with the value of the threshold, and that for at least one value of the threshold (0.95) COMB outperforms both TTC and BASE by 0.20 BLEU points. To confirm this finding with another connective, we took the first 1,572 sentences containing the connective *since* from the same UN test set. The BLEU scores for COMB are shown for thresholds within the interval of observed confidence values (0.4–1.0) in Figure 7.3(b). For several values of the threshold, COMB outperforms both BASE and TTC, in particular for 0.85, with a difference of 0.39 BLEU points.



(a) *although* (b) *since*

Figure 7.3: Use of a combined system (COMB) that directs the input sentences either to a system trained on a sense-labeled corpus (TTC) or to a baseline one (BASE), depending on the confidence of the connective classifier. The x-axis shows the threshold above which TTC is used and the y-axis shows the BLEU scores of COMB, TTC and BASE.

The significance of the observed improvement of COMB versus TTC and BASE was tested as follows. For each of the two connectives, we split each of the test sets of 1,572 sentences into five folds, and compared for each fold the scores of COMB for the best performing threshold (0.95 or 0.85) with the highest of BASE or TTC (i.e. BASE for *although* and TTC for *since*). We performed a paired t-test to compute the significance of the difference, and found $p = 0.12$ for *although*. This value, although slightly above the conventional boundary of 0.1, shows that the five pairs of scores reflect a notable difference in quality. When performing the t-test for *since*, the difference in scores was found to be statistically significant at the 0.01 level ($p = 0.005$). Moreover, COMB was found to be always significantly better than the lower of BASE or TTC

($p < 0.05$). This experiment was reported in (Meyer and Popescu-Belis [2012]).

Although this method showed improvements in translation quality for discourse connectives, the system design is relatively complex (because both a baseline and a discourse-aware SMT system are needed) and the thresholds that determine which systems to use are set empirically. Below, we describe an alternative experiment also using the confidence scores output by the classifier, but including this information directly into the data for SMT training and tuning by duplicating it.

7.5 Duplication of training data based on label confidence

To incorporate the maximum of information from the discourse connective classifiers, we experimented, in MT training and tuning, with the label probability distribution (or confidence score) for each connective, obtained from the MaxEnt models. Let us consider the following example:

Last year, people 60 and older accounted for almost 22 percent of Shanghai's registered residents, ***while*** *the birthrate was less than one child per couple.*

For the EN connective *while*, the automatic classifier found that it signals most probably a CONTRAST discourse relation ($p = 0.67$), but it might also signal a CONCESSION ($p = 0.29$), along with other less likely possibilities (TEMPORAL-CONTRAST and TEMPORAL). In total, for the six connectives considered here, there are 12 possible sense labels.

In the present experiment, the labels and their scores, as output by the classifier in Table 5.7, were used for all six connective types mentioned there and labels were assigned by label-concatenation (Section 7.3). To model the label probability distributions directly in the training and tuning phases of SMT systems, we generate in the training data ten copies of each labeled sentence, and label each of them according to the discretized probability distribution with 10 bins (from 0 to 1 with 0.1 increments). In the example above, we produce 7 copies of the sentence with the connective labeled CONTRAST and 3 copies with the label CONCESSION. All unlabeled sentences are also copied 10 times to keep the original proportions in the data. In this way, the occurrences of labels seen by the SMT system are a reflection of the confidence of the classifier in the label decisions. The counterpart baseline SMT system is also trained on the same, multiplied amount of data, but without any labels. The same procedure is applied to the data used for MT tuning.

The systems were built using tokenized and lowercased data, with the Moses SMT toolkit (Koehn et al. [2007]) and were tuned using MERT (Och [2003]). The following data sets were used.

For training, we used Europarl v6 for EN/FR (direct translations only), 321,577 sentences, with 9,038 occurrences of the six connectives labeled automatically. For tuning, we used the News Commentary 2011 tuning set (3,003 sentences), with 133 occurrences labeled automatically.

For testing, we used the WMT 2010 shared translation task test data (2,489 sentences), with 140 occurrences labeled automatically. The language model was a 3-gram one over a combination of all French texts of Europarl and News Commentary, built by using the IRSTLM toolkit (Federico et al. [2008]). For testing, we only input to SMT the most probable label output by the classifier.

The results of the system trained using the duplicated data following the label probabilities (noted LPD) and those of the baseline system are given in Table 7.7. Along with BLEU, we report the ACT scores for the modified and the baseline SMT systems (see Section 4.3). Besides its automatic variants (ACT_a and ACT_{a5+6}), we also manually looked through the cases 5 and 6 of ACT to spot actually correct translations and report the most precise ACT score (ACT_m) in Table 7.7. The ACT scores are quite similar for the LPD system and the baseline, therefore this method of multiplying the data based on classifier label confidence does not seem to help much in terms of connective translation quality. The overall BLEU score however, is improved by 0.3 points. While this variation could be due to differences between MERT tuning runs (Section 4.3), it still shows that changing the labels of discourse connectives at least clearly preserves the global performance measured by BLEU.

Translation model	**SMT system**	**BLEU**	ACT_a	ACT_{a5+6}	ACT_m
Phrase-based with label probabilities	LPD	21.60	69.4	82.0	78.5
	Baseline	21.30	68.8	81.1	79.2

Table 7.7: BLEU and ACT scores on WMT10 test data for a system using duplicated training data in proportion of the probability distributions of the labels on discourse connectives.

Besides ACT and BLEU, we compared the connective translations by the LPD system to the ones output by the baseline in terms of Δ *Connectives*. We obtain very similar scores to the ones given in Section 7.3 above: about 11% of the connectives are improved, while 85% remain the same and 4% are degraded by the modified system.

To estimate the maximal improvement of BLEU if all connectives were translated correctly, we considered the WMT10 test set. In the output generated the LPD model, we changed, where necessary, the occurrences of the discourse connectives to make them *identical* to the human reference translation, without changing any other word. As a result, 73 occurrences were altered, leading to an improvement of the BLEU score of 0.17 points (to 21.77 vs. 21.60 for LPD in Table 7.7). We then performed similar changes to the baseline system output. Nearly the same number of connectives (70) were altered, leading to a similar improvement in BLEU of 0.18 (to 21.48 compared to 21.30 as shown for the baseline in Table 7.7).

This shows that even if all connectives are translated as in the reference, the improvement of BLEU remains, as expected, very moderate. Of course, the similar number of changes does not reflect the quality of the LPD system, which generates more correct connectives than the baseline, although not identical to the reference ones, as can be seen from the detailed ACT categories.

In another test, we computed the BLEU score separately for each segment (i.e. sentence) that contains one of the six targeted connectives, and then compared the scores for each pair of segments from the baseline and the LPD system. We found that 29% of the segments had a higher BLEU score when translated by LPD than by the baseline system, 55% had the same scores for LPD and baseline, and 16% had a lower score when translated by LPD than by the baseline system. Again, this demonstrates the improvement brought by LPD. The above experiments and evaluations were part of (Meyer et al. [2012]).

7.6 Post-editing discourse connectives

The ACT metric (see Hajlaoui and Popescu-Belis [2013] and Section 4.3 of this thesis) incorporates heuristics for word alignment applied to connectives, along with lists of acceptable translations of connectives depending on their identified senses. These can be used to post-edit the output of SMT in order to correct target connectives that are incompatible with the sense signaled by the source connective, as found by our automatic connective labeler. For instance, in the example shown in Figure 2.1 in Chapter 2, if the source connective *since* is labeled as TEMPORAL, and an MT system generates the French causal connective *parce que*, then through post-editing this can be corrected to one of the acceptable temporal French translations of *since*, for instance *depuis que*.

We have experimented with the output of the SMT systems for EN/FR and EN/DE which are described below in Section 7.7.2, including the same tuning, but with the difference that all data was lowercased. The six targeted connectives were labeled by the `All_Features` model – its results are shown in Table 5.10 of Section 5.6 above.

Comparing the baseline EN/FR SMT with the post-edited output on the nt2012 data set, the BLEU scores were identical (26.7), while ACT scores were respectively 56.28 and 56.48 when averaged over 5 MERT tuning runs. Although slightly higher, the score of the post-edited version is not significantly different from the baseline. For EN/DE, the BLEU scores are nearly identical (12.0 vs. 11.9) while ACT scores increased from 62.28 to 65.58, which is a significant improvement ($p < 0.001$). An explanation of the difference between EN/FR and EN/DE is that in the set of sentences that were actually post-edited (31 for FR and 37 for DE, out of 176 occurrences in nt2012), there were more correct connective labels in the EN/DE data than in the EN/FR dara (25 vs. 13). This suggests that post-editing is a viable strategy if label accuracy was improved. Indeed, we also scored a post-edited output with *oracle* labels, with ACT scores of 59.58 for EN/FR and 66.66 for EN/DE, which were both significantly above the baseline ($p < 0.001$).

The manual scoring of the post-edited output, performed on a 1-to-4 scale by three FR (respectively DE) native speakers, showed however that for EN/FR it is the baseline translations that were considered significantly better than the post-edited ones (2.5 vs. 2.0, $p < 0.05$), as well as for EN/DE (3.2 vs. 2.5, $p < 0.01$). Therefore, this post-editing strategy (presented in (Meyer et al. [2014])) appears to produce results that are less acceptable to human judges,

but similar or even better in terms of BLEU and ACT. The approach was not further pursued, though it could yield better results when the automatic disambiguation of the connectives would further be improved. The next section presents factored translation models, with which we achieved the best improvements for connective translation, and which also represent the most principled solution to integrate linguistic labels into SMT among those studied in this thesis.

7.7 Factored Models

Factored translation models (Koehn and Hoang [2007]) for phrase-based SMT systems offer a principled way to use linguistic labels (e.g. morpho-syntactic, semantic, or discourse ones) and do not require human intervention in the data or phrase tables. Such models have most often been used to integrate morphological information, for instance when translating into a morphologically rich language. But, as we will show in the experiments below, they also lead to the highest improvements in terms of connectives, when compared to the models described above.

Phrase-based factored translation models combine features in a log-linear way, as shown in the equation below for the most probable target sentence $\hat{f}$ to be found when decoding. In this equation, M is the number of features, $h_m(e_1^{F_e}, f_1^{F_f})$ are the feature functions over the factors, and λ_m are the weights for combining the features, which are optimized during MERT tuning. Each feature function depends on a vector $e_1^{F_e}$ (in our case e_{wl} for source words and labels) and a vector $f_1^{F_f}$ (in our case f_w for target words).

$$\hat{f} = \arg\max_{f} \left\{ \sum_{m=1}^{M} \lambda_m \cdot h_m(e_1^{F_e}, f_1^{F_f}) \right\}$$

Although both source and target factors can be used, we consider source-side factors only, as our annotation of discourse relations is done on the source. We will combine, on the source side, discourse labels on connectives with part-of-speech tags on all words.

Figure 7.4 shows an example sentence, where instead of plain text (sentence 1) as input for the SMT system one augments words with labels: part-of-speech (POS) tags (sentence 2), POS tags combined with discourse labels (DL) for connectives (3), or discourse labels only (4), in which case all other labels are set to null. In our experiments, the POS tags were generated by the Stanford POS tagger (Toutanova et al. [2003]) with the `bidirectional-distsim-WSJ` model.

For building the translation models and for MERT tuning, both the English source word and the factor information are used to generate the surface target language word forms. We designed three different MT systems (in addition to the baseline one), using either `POS` factors, or `POS+DL`

1. for the first time it was said that the countries who want are to cooperate, **while** those who are not willing can stand off.

2. for|in the|dt first|jj time|nn it|prp was|vbd said|vbd that|in the|dt countries|nns who|wp want|vbp are|vbp to|to cooperate|vb ,|, **while**|**in** those|dt who|wp are|vbp not|rb willing|jj can|md stand|vb off|rp .|.

3. for|in the|dt first|jj time|nn it|prp was|vbd said|vbd that|in the|dt countries|nns who|wp want|vbp are|vbp to|to cooperate|vb ,|, **while**|**in-contrast** those|dt who|wp are|vbp not|rb willing|jj can|md stand|vb off|rp .|.

4. for|null the|null first|null time|null it|null was|null said|null that|null the|null countries|null who|null want|null are|null to|null cooperate|null ,|null **while**|**contrast** those|null who|null are|null not|null willing|null can|null stand|null off|null .|null

Figure 7.4: Example sentence for factored translation models: (1) plain text, (2) POS tags as factors, (3) POS tags combined with discourse labels (DL), and (4) DL only.

factors, or only DL factors. All data (training, tuning and test) has to be factored in the same way for each system. We built the factored translation models using the labels which were output by our classifiers of discourse connectives (see Chapter 5), which had been previously trained on Europarl and PDTB data. This approach (as opposed to using manually annotated data) offers a large (but noisy) data set for MT training, tuning and testing, limited only by the amount of parallel data available. In addition, since labels are modeled as additional features, the method does not suffer from sparsity as the one using label-concatenation (Section 7.3). Of course, labels are not always correct, as the performance of the connective classifiers is in the range of 0.7–1.0 F1-score.

7.7.1 Factored models with discourse and POS labels

We first built factored translation models over Europarl v6 EN/FR with the exact same training/tuning/test and language model data described for the LPD model above (Section 7.5).

We built a factored phrase-based model with the Moses SMT toolkit (Koehn et al. [2007]) and compared it to a factored *hierarchical* phrase-based model built with the cdec decoder (Dyer et al. [2010], see also Chapter 3). The latter was to see whether including hierarchical (syntactical) features in the translation process would change the translation quality for discourse connectives (which it does not, as we show in the following).

Labels were assigned to six discourse connectives (*although, meanwhile, since, (even) though, while, yet*) automatically in all data by the classifier listed in Table 5.7. The baseline systems were built over texts only, not considering any labels or factors. The results, in terms of BLEU and ACT scores, of the phrase-based models with POS tags and/or POS tags and discourse labels (DL), and of the hierarchical model with discourse labels are presented in Table 7.8.

The phrase-based factored systems clearly outperform the plain text phrase-based baseline in terms of the correct translation of the connectives, and using combined factors (POS+DL) brings the highest improvement. For ACT_m, which gives the most precise assessment of this improvement, POS+DL achieves the highest scores, as it translates 1.4% of the connectives better than DL alone (absolute difference), 5.7% better than POS and 8.5% better than the plain

Translation model	SMT system	BLEU	ACT_a	ACT_{a5+6}	ACT_m
Factored phrase-based	POS + DL	22.19	70.7	86.1	82.1
	DL	21.69	70.0	85.2	80.7
	POS	22.26	67.9	81.2	76.4
	Baseline	21.71	65.0	77.8	73.6
Factored hierarchical	DL	19.20	67.9	78.5	77.1
	Baseline	19.31	63.6	74.8	74.3

Table 7.8: BLEU and ACT scores on WMT10 for translation models that use automatically labeled connectives and two baseline ones. Source-side factors are part-of-speech tags, used alone (POS) or in combination with labeled connectives (POS+DL), and discourse labels only (DL). The ACT scores are highest for the phrase-based factored model using both POS and DL.

text baseline. These scores also tend to show that, as expected, the factoring of discourse connective labels brings more improvement than the use of POS (compare DL/baseline with POS/baseline: +7.1% vs. +2.8% absolute). The other versions of the ACT score vary in the same direction as ACT_m and confirm these findings. For the hierarchical factored model, the experiments show that the DL system translates 2.8% of the connectives above baseline, in terms of ACT_m.

It is also possible to estimate the effect of the factors in terms of improved / unchanged / degraded connectives in the translations of a modified system compared to a baseline. When counted over the WMT10 set for the POS+DL system, about 16% of the connectives are improved with respect to the baseline, 81% are unchanged, and only 3% are degraded. When counting the same for the hierarchical factored translation model with discourse labels, 11% of the connective translations are improved, 86% remained unchanged, and 3% were degraded. These scores are slightly superior to those we obtained using a concatenated connective-label model instead of factors (Section 7.3). The improvements for connective translation with those models were in a range of 11 to 18%, with 60–70% unchanged connectives, and a higher number of degraded translations (14–24%).

For the BLEU scores, as expected, variation is quite small, since the number of changed words with respect to the reference is small (see also the estimates with 100% correct connectives at the end of Section 7.5 above). Still, our phrase-based factored models show an improvement in BLEU with respect to the baseline, but this seems mainly due to the POS factors: +0.48 for POS+DL and +0.55 for POS. The use of the discourse labels only (DL) leaves BLEU almost unchanged, or decreases it very slightly as in the case of the hierarchical factored model. It is also possible that these variations are due to the different runs of the MERT tuning. These findings were presented in (Meyer et al. [2012]).

Over time we extended our classifier models with more features (namely dependency, polarity, discursive – see Section 5.6). In the following sub-section, we therefore present experiments for factored phrase-based SMT that operate on the entire Europarl corpora (not only the direct

translations), applied to several target languages, with more connective types and the best classifier models trained jointly on Europarl and PDTB data. We will not further consider hierarchical models due to their slightly lower performance.

Neither will we consider models incorporating POS tags, as we mainly want to measure improvements to translation quality caused by discourse labels (DL) on connectives only.

7.7.2 Factored models with discourse labels across multiple target languages

Using the Moses decoder (Koehn et al. [2007]), we built MT systems from English to four target languages: French, German, Italian, and Arabic. The baseline systems were built without any modification to the text from the corpora, except for tokenization and truecasing with the Moses tools. The language models were 3-gram ones built by using the IRSTLM toolkit (Federico et al. [2008]). For Italian, they were built using Europarl v7, while for French and German they were built over a combination of Europarl v7 and the News Commentary corpus (years 2007-2011), as distributed by the Workshops on statistical MT. For Arabic, we built a 3-gram language model from the UN corpus. Optimization was done using MERT (Och [2003]) as provided with Moses.

The labeling of now seven EN discourse connectives (*although, however, meanwhile, since, (even) though, while, yet*) was done by using the MaxEnt classifier and the `All_features` model (Table 5.10, Section 5.6) after having found that it gives the most robust scores on unseen text. The BLEU scores are reported using the MultEval tool (see Section 4.3) on tokenized and truecased text with the Moses tools. All reported system scores are averaged over five runs of MERT tuning, in order to mitigate its non-deterministic approach.

Data

The data for the experiments was chosen according to established practice, aiming for testing sets of similar sizes. Table 7.9 shows the data sets, in terms of origins, genre, numbers of sentences and of labeled connectives that we used for building and testing our SMT systems. The data for EN/FR, EN/DE and EN/IT is distributed by the WMT workshop [7]. Data pre-processing for these three language pairs consisted of tokenization and truecasing. For EN/AR the data is licensed from the United Nations Corpora [8] and from the Linguistic Data Consortium for the NIST OpenMT evaluation sets [9]. The English side was again tokenized and lowercased, while Arabic was transliterated and words were segmented using MADA (Habash and Rambow [2005]).

System tuning and testing is performed over news articles with a variety of topics. While the EN/FR and EN/DE systems were tuned and tested on the data sets with the same EN source,

7. http://www.statmt.org/wmt12/translation-task.html
8. http://www.uncorpora.org/
9. http://catalog.ldc.upenn.edu/LDC2013T03

Language pair	Role	Data source	Genre	# Sentences	# DL
EN/FR	training	EP	parl. debates	1,998,684	139,585
	tuning (1)	nt2011	newswire	3,003	174
	testing (2)	nt2012	newswire	3,003	176
	testing (3)	nt2010	newswire	2,489	165
	testing (4)	nt2008 and sy2009	newswire	2,502	122
EN/DE	training	EP	parl. debates	1,906,486	133,448
	tuning (1)	nt2011	newswire	3,003	174
	testing (2)	nt2012	newswire	3,003	176
	testing (3)	nt2010	newswire	2,489	165
	testing (4)	nt2008 and sy2009	newswire	2,502	122
EN/IT	training	EP	parl. debates	1,898,118	138,381
	tuning	nt2009	newswire	2,525	201
	testing (4)	nt2008 and sy2009	newswire	2,502	122
EN/AR	training	UN	parl. debates	5,989,646	242,248
	tuning	nist2006 and nist2008 and nist2009	newswire & web	6,099	347
	testing	nist2002 to nist2005	newswire & web	3,522	176

Table 7.9: Genres, sizes and numbers of (labeled) connectives (DL) in the data for training, tuning and testing SMT systems. The data sets are: EP (Europarl corpus v. 7), nt (newstest), sy (newssyscomb), UN (United Nations corpus), nist (NIST OpenMT). Identical numbers in parentheses indicate identical source sides.

this was not the case for EN/IT and EN/AR. However, one test set is shared across EN/FR, EN/DE, and EN/IT.

The performance of SMT systems is sensitive to the similarity between the training/tuning and the test data. With the MERT tuning method (Och [2003]) it is emphasized that tuning data should be from the same domain and genre as the test data, for tuning to improve output quality. We examined the similarity between the EN sides of our data sets, using cosine text similarity from the software implemented by Pedersen et al. (v0.10, June 2013) [10].

Overall, the similarity of the testing sets for FR-DE-IT with the respective tuning sets is around 0.74–0.78, but this value is markedly lower for AR, at only 0.64. The similarity of the test sets with the training sets is even lower, around 0.50–0.55 for all four languages. The similarities between the three test sets used for EN/FR and EN/DE (noted (2)-(3)-(4) in Table 7.9) are in the same range (0.74–0.77). However, the distribution of the seven EN connective types differs quite markedly across these three sets, as shown in Table 5.13 on page 73 above. For instance, the proportion of *since* varies between 17% and 37%, and that of *while* between 9% and 34%.

Results

The BLEU and ACT scores obtained for the four target languages and four test sets are shown in Table 7.10. The significance values (coded as '*', '**' or '***') of the differences between the baseline SMT system and the SMT systems with labeled connectives were obtained over five independent runs of the MERT tuning algorithm. The scores vary considerably depending on the training and testing sets and the language pair, and our main goal now is to assess the

10. http://text-similarity.sourceforge.net/. The cosine similarity scores (between 0 and 1) were computed over term-frequency vectors, from lowercased texts excluding punctuation.

Languages	Test set	System	BLEU	Δ	p	ACT	Δ	p
EN/FR	nt2012	baseline	26.1			56.28		
		DL	25.8	-0.3	**	57.68	1.40	*
	nt2010	baseline	24.4			68.12		
		DL	24.3	-0.1	**	68.60	0.48	*
	nt2008+sy2009	baseline	28.9			61.36		
		DL	29.2	0.3	*	60.94	-0.42	*
EN/DE	nt2012	baseline	11.8			62.28		
		DL	11.8	0.0	n/s	65.08	2.80	**
	nt2010	baseline	15.0			62.42		
		DL	15.0	0.0	n/s	69.28	6.86	***
	nt2008+sy2009	baseline	13.0			71.06		
		DL	13.1	0.1	n/s	70.30	-0.76	n/s
EN/IT	nt2008+sy2009	baseline	23.7			77.10		
		DL	24.1	0.4	*	76.78	-0.32	n/s
EN/AR	nist2002–nist2005	baseline	18.2			64.72		
		DL	18.3	0.1	*	62.20	-2.52	*

Table 7.10: BLEU and ACT scores averaged over five optimizer runs. For each language pair and test set, we indicate the score difference (Δ) between the baseline SMT system and the system that uses as source-side factors the automatically-assigned discourse connective labels (DL). The statistical significance of the difference (p-value of paired t-test over the five runs) is noted with * for the 10% level, ** for the 1% level and *** for 0.1% level (most reliable difference).

improvement brought by labeled connectives in each condition.

As expected, given the sparsity of connectives, the BLEU scores do not necessarily increase when using labeled connectives; neither do they decrease considerably. The BLEU scores increase slightly (but with statistical significance) for EN/DE and EN/IT when testing on nt2008+sy2009, as well as for EN/AR when testing on nist2002–nist2005. However, they decrease slightly (again with statistical significance) for EN/FR on nt2010 and nt2012. Our conclusion, as with all models above, is that the use of labeled connectives does not degrade single-reference BLEU scores.

Turning now to the discourse connectives, most of the variations of the ACT metric indicate a significant improvement in the translation of connectives when using our solution for the EN/FR and EN/DE systems and the nt2010 and nt2012 data sets (up to 7 ACT points). This validates our proposal as a viable method to improve the translation of connectives through their separate automatic labeling.

However, the negative results in Table 7.10 must also be understood. The lack of improvement when using labeled connectives is apparent when testing on the nt2008+sy2009 data, for EN/FR, EN/DE and EN/IT alike. When examining this data set in terms of genre, topics, or even cosine similarity, no marked difference is found with nt2010 or nt2012. However, as shown in Section 5.6.4, Table 5.13, the accuracy of connective labeling on nt2008+sy2009

is lower (F1 = 0.61) than on nt2010 (F1 = 0.64) and especially nt2012 (F1 = 0.72), due to the different proportions of easy vs. difficult connectives. These differences are reflected in the ACT improvements (Δ), or lack thereof, on the different test sets, and explain in particular the lack of improvement for all the target languages on nt2008+sy2009 – a data set on which connective labeling is insufficiently accurate. If labeling for the difficult connectives would be improved beyond a certain threshold (appearing, in our data, to be at around 0.70 F1), their translation when using discourse-aware MT would become more accurate, as in the case of nt2010 and nt2012.

In the case of EN/AR, the ACT score on nist2005–nist2009 is degraded most compared to the other language pairs. Upon manual inspection of the labels output by our classifier we also noticed its lower accuracy. This is due to the difference of this data (web+newswire) to EP+PDTB (debates+newswire).

For the earlier factored models (Section 7.7.1), the ACT score on nt2010 for EN/FR improved by up to 5.7 points, which is higher than the improvement shown in Table 7.10 (0.48 points). We here made use of all Europarl data available for EN/FR, whereas in Section 7.7.1, only the original EN and direct FR translations of the EN/FR pair in Europarl have been used. With such reduced data, discourse-aware MT contributed more noticeably to improve connective translation. In the experiments presented in this section, however, due to more training data, the baseline system reaches a higher translation quality which is confirmed by its higher BLEU score (24.4 for EN/FR on nt2010 vs. 21.7 in Section 7.7.1 on the same test set). These experiments were presented in (Meyer et al. [2014]).

7.8 SMT with labels for verb tense

Along the lines of the above experiments for discourse connectives, we applied similar methods to explore verb tense translation in SMT. The goal was two-fold. On the one hand, we would like to show the generalizability of label concatenation and factored translation models to other discursive phenomena. On the other hand, we wanted to see whether pairing a classifier to predict verb tense with an SMT system leads to translation improvements as well. We tested SMT systems that rely on labels output by the narrativity classifier (described in Section 6.1), in Section 7.8.1 below, and with labels by the French tense predictor (described in Section 6.2), in Section 7.8.2 below. Based on the results on connectives, factored translation models were selected for all the experiments with verb tenses.

7.8.1 SMT with narrativity labels

Two methods to convey information about narrativity to an SMT system were explored. First, as in our initial studies applied to discourse connectives, the narrativity labels were simply concatenated with the Simple Past (SP) verb form in EN as in the second line of Figure 7.5. Second, we used factored translation models, to combine tense labels on verbs with the

basic features of phrase-based SMT models (phrase translation, lexical and language model probabilities).

To assess the performance gain of narrativity-augmented systems, we built three different SMT systems, with the following names and configurations:

BASELINE: plain text, no verbal labels.

TAGGED: plain text, all SP verb forms concatenated with a narrativity label.

FACTORED: all SP verbs have narrativity labels as source-side translation factors (all other words labeled 'null').

1. BASELINE SMT: on wednesday the čssd declared the approval of next year's budget to be a success. the people's party was also satisfied.
2. TAGGED SMT: on wednesday the čssd declared-**Narrative** the approval of next year's budget to be a success. the people's party was-**Non-narrative** also satisfied.
3. FACTORED SMT: on wednesday the čssd declared|**Narrative** the approval of next year's budget to be a success. the people's party was|**Non-narrative** also satisfied.

Figure 7.5: Example input sentence from 'nt2010' data for three translation models: (1) plain text; (2) concatenated narrativity labels; (3) narrativity as translation factors (the '|null' factors on other words were omitted for readability).

Figure 7.5 shows an example input sentence for these configurations. For the FACTORED SMT model, both the EN source word and the factor information are used to generate the FR surface target word forms. The tagged or factored annotations are respectively used for the training, tuning and testing. For labeling the SMT data, no manual annotation is used. In a first step, the actual EN SP verbs to be labeled are identified using the Stanford POS tagger (Toutanova et al. [2003]), which assigns a 'VBD' tag to each SP verb. These tags are replaced, after feature extraction and execution of the MaxEnt classifier, by the narrativity labels output by the latter (as explained in Section 6.1 above). Of course, the POS tagger and (especially) our narrativity classifier may generate erroneous labels, which in the end lead to translation errors. The challenge is thus to test the improvement of SMT with respect to the baseline, in spite of the noisy training and test data.

Data

In all experiments, we made use of parallel English/French training, tuning and testing data from the translation task of the Workshop on Machine Translation (www.statmt.org/wmt12/). For *training*, we used Europarl v6 (Koehn [2005]), original EN to translated FR (321,577 sentences), with 66,143 instances of SP verbs labeled automatically: 30,452 are narrative and 35,691 are non-narrative. For *tuning*, we used the Newstest 2011 tuning set (nt2011, 3,003 sentences), with 1,401 automatically labeled SP verbs, of which 807 are narrative and 594

non-narrative. For *testing*, we used the Newstest 2010 data (nt2010, 2,489 sentences), with 1,156 automatically labeled SP verbs (621 narrative and 535 non-narrative).

We built a 5-gram language model with SRILM (Stolcke et al. [2011]) over the entire FR part of Europarl. Tuning was performed with MERT. All translation models were phrase-based using either plain text (possibly with concatenated labels) or factored training as implemented in the Moses SMT toolkit. All data was tokenized and lowercased using the Moses tools.

Results: automatic evaluation

In order to obtain reliable automatic evaluation scores, we executed three runs of MERT tuning for each translation model. Table 7.11 shows the average BLEU scores on the nt2010 data for the three systems. The scores are averages over the three tuning runs, with resampling of the test set, both provided in the MultEval evaluation tool (Section 4.3). A t-test was used to compute p values that indicate the significance of differences in scores.

Translation model	BLEU
BASELINE	21.4
TAGGED	21.3
FACTORED	21.6*

Table 7.11: Average values of BLEU scores over three tuning runs for each model for verb translation on nt2010. The starred value is significantly better ($p < 0.05$) than the baseline.

The FACTORED model improves performance over the BASELINE by +0.2 BLEU, a difference that is statistically significant at the 95% level. On the contrary, the concatenated-label model (noted TAGGED) slightly decreases the global translation performance compared to the BASELINE. A similar behavior was observed when using labeled connectives in combination with SMT (Section 7.3).

The lower scores of the TAGGED model may be due to the scarcity of data (by a factor of 0.5) when verb word-forms are altered by concatenating them with the narrativity labels. The small improvement of the FACTORED model may also be related to the scarcity of SP verbs: although their translation is definitely improved, as we will now show, the translation of all other words is not changed by our method, so only a small fraction of the words in the test data are changed with respect to the baseline.

Results: human evaluation

To assess the improvement that is specifically due to the narrativity labels, we manually evaluated the FR translations by the FACTORED model for the first two hundred SP verbs in the test set against the translations from the BASELINE model. As the TAGGED model had lower BLEU scores and appeared to translate verb phrases less accurately upon informal inspection,

we did not submit it to human evaluation. Manual scoring was performed along the following criteria for each occurrence of an SP verb, by bilingual judges looking both at the source sentence and its reference translation.

— Is the narrativity label correct? ('correct' or 'incorrect', a direct evaluation of the narrativity classifier from Section 6.1)
— Is the verb tense of the FACTORED model more accurate than the BASELINE one? (noted here with '+' if improved, '=' if similar, '−' if degraded)
— Is the lexical choice of the FACTORED model more accurate than the BASELINE one, regardless of the tense? (again noted '+' or '=' or '−')
— Is the BASELINE translation of the verb phrase globally correct? ('correct' or 'incorrect')
— Is the FACTORED translation of the verb phrase globally correct? ('correct' or 'incorrect')

Tables 7.12 and 7.13 summarize the counts and percentages of improvements and/or degradations of translation quality of the FACTORED system vs. the BASELINE one. The correctness of the labels, as evaluated by the human judges on SMT test data, is similar to the values given in Section 6.1 when evaluated against the test sentences of the narrativity classifier. As shown in Table 7.12, the narrativity information clearly helps the FACTORED system to generate more accurate French verb tenses in almost 10% of the cases, and also helps to find more accurate vocabulary for verbs in 3.4% of the cases. Overall, as shown in Table 7.13, the FACTORED model yields more correct translations of the verb phrases than the BASELINE in 9% of the cases – a small but non-negligible improvement.

Criterion	**Rating**	**N.**	**%**	Δ
Labeling	correct	147	71.0	
	incorrect	60	29.0	
Verb tense	+	35	17.0	+9.7
	=	157	75.8	
	−	15	7.2	
Lexical choice	+	19	9.2	+3.4
	=	176	85.0	
	−	12	5.8	

Table 7.12: Human evaluation of verb translations into French, comparing the FACTORED model against the BASELINE. The Δ values show the clear improvement of the narrativity-aware factored translation model.

An example from the test data shown in Figure 7.6 illustrates the improved verb translation. The BASELINE system translates the two EN SP verbs *worked* and *was* incorrectly in French with a participle only (*travaillé*) and no verb at all, respectively. The FACTORED model generates the correct and complete tense (PC, *a travaillé*) and the verb *est* in the second case (which should however be in IMP tense, as *était* in the reference translation). The first sentence is scored as follows: the labeling is correct ('yes'), the tense was improved ('+'), the lexical choice was the same ('='), the BASELINE was incorrect while the FACTORED model was correct. The second

System	Rating	Number	%
BASELINE	correct	94	45.5
	incorrect	113	54.5
FACTORED	correct	113	54.5
	incorrect	94	45.5

Table 7.13: Human evaluation of the global correctness of 207 translations of EN SP verbs into French. The FACTORED model yields 9% more correct translations than the BASELINE one.

sentence is scored as labeling correct ('yes'), the tense was the same ('='), the lexical choice was improved ('+') and both, the BASELINE and the FACTORED model were incorrect.

EN: freeman **worked**|**Non-narrative** for several years to get mandela 's story onto the big screen ... the most important thing **was**|**Non-narrative** that he wanted to shake mandela 's hand .
FR REFERENCE: freeman **a travaillé** quelques années pour amener l' histoire de mandela sur grand écran ... le plus important pour lui **était** de lui serrer la main .
FR BASELINE: freeman **travaillé** pendant plusieurs années à obtenir mandela sur le grand écran de l' histoire ... la chose la plus importante **__0__** qu' il voulait mandela serrer la main .
FR FACTORED: freeman **a travaillé** pendant plusieurs années à la grande histoire de mandela à l' écran ... la chose la plus importante **est** qu' il voulait mandela serrer la main .

Figure 7.6: Example comparison of a baseline and improved factored translation.

When looking in detail through the translations that were degraded by the FACTORED model, some were due to the POS tagging used to find the EN SP verbs to label. For verb phrases made of an auxiliary verb in SP and a past participle (e.g. *was born*), the POS tagger outputs *was*/VBD *born*/VBN. As a consequence, our classifier only considers *was*, as non-narrative, although *was born* as a whole is a narrative event. This can then result in wrong FR tense translations. For instance, the fragment *nelson mandela was*|Non-narrative *born on* ... is translated as: *nelson mandela *était né en* ..., which in FR is a Plus-que-parfait (pluperfect) tense instead of the correct Passé Composé *est né* as in the reference translation. These findings for SMT with the narrativity feature were published in (Meyer et al. [2013]).

In an alternative approach, illustrated in the next section, a more direct way to label verb tenses is to use a classifier with similar features, but using as classes the desired target verb tenses, although they cannot always be accurately predicted.

7.8.2 SMT with predicted French tense labels

In another series of experiments, we considered all verbs regardless of their tense, without using the intermediate category of narrativity, which is difficult to predict. We used the classifiers described in Section 6.2 to predict the FR tense label to which the EN verbs should be translated, prior to the training of a factored translation model. This method has the advantage of providing much more training data, which is extracted from the alignment of the verb phrases, as explained in Section 4.2.2. It generality makes it applicable to all tenses, not only SP. Moreover, this method is likely to learn which verbs are preferably translated with which tense: for instance, the verb *started* is much more likely to become *a commencé* (PC) in FR than to *commençait* (IMP), due to its meaning of a punctual event in time, rather than a continuous or repetitive one.

We compare the three models of FR tense prediction described in Section 6.2 in terms of their effect on SMT; moreover, we also compare them against the oracle experiment from Section 7.1.2 above. The data used for building factored phrase-based SMT models was exactly the same as in Section 7.1.2, i.e. 203,140 sentences from the EN/FR Europarl corpus v7, from which 7000 sentences were subtracted: 4000 for tuning and 3000 for testing. The important difference with the oracle experiment is test set no longer uses gold-standard labels, but those that are output by one of the three prediction models. Therefore, the same translation model as in the oracle experiments can be used – namely a factored phrased-based Moses model, trained and tuned with verbs with gold labels, and a 3-gram IRSTLM language model over Europarl v7 FR plus the FR side of the News Commentary corpus (years 2007–2011). All scores will again be averaged over 3 runs of MERT in order to account for stability. The baseline system uses the same data, without considering factors or labels.

The approach is comparable to the setting of the experiment 1 with connectives described in Section 7.3. However, thanks to the automatic alignment of verb phrases, which generates the gold-standard set, we can rely on many more labeled instances in the SMT training and tuning data (i.e. hundreds of thousands), unlike the case of discourse connectives. We thus only labeled the test automatically (since we aim for fully automatic MT) with each of the three tense prediction models defined in Section 6.2.

Configuration	**BLEU**	Δ Base	**10-fold c.v.**	**Test set**
Baseline	27.67	–	–	–
Oracle	28.17	**0.50**	–	–
ALL-CLASSES	27.72	0.05	0.75	–
9-CLASSES	27.78	**0.11**	0.85	0.83
EXTENDED	27.79	0.12	0.85	0.83

Table 7.14: BLEU scores, difference with baseline BLEU scores, and verb tense classifier performance (10-fold cross-validation on the training set, then scores on the test set) for five configurations of the SMT system.

.

Tense	Baseline	Oracle	Δ Base	9-CL.	Δ Base	# sent.	10-fold c.v.	Test set
Imparfait	24.10	25.32	**1.22**	24.41	0.31	122	0.475	0.400
Passé Composé	29.80	30.82	**1.02**	30.07	0.27	636	0.769	0.726
Impératif	19.08	19.72	0.64	18.07	-1.01	4	0.286	0.000
Passé Simple	13.34	16.15	**2.81**	14.02	0.68	6	0.158	0.000
Plus-que-Parfait	21.27	23.44	**2.17**	22.07	0.80	17	0.547	0.361
Présent	27.55	27.97	0.42	27.59	0.04	2618	0.918	0.905
Subjonctif	26.81	27.72	0.91	26.11	-0.70	78	0.329	0.155
Passé Récent	24.54	30.50	**5.96**	26.56	**2.02**	3	0.162	0.000
Average/Total	23.31	25.21	**1.89**	23.61	0.30	3484	0.456	0.318

Table 7.15: BLEU and F1 scores of the classifier, for each predicted FR tense for three systems: baseline SMT, oracle SMT and SMT with tense predictions from the model 9-CLASSES (9-Cl.).

Tense	Baseline	Oracle	Δ Base	EXT.	Δ Base	# sent.	10-fold c.v.	Test set
Imparfait	24.10	25.32	1.22	24.57	0.47	122	0.47	0.44
Passé Composé	29.80	30.82	1.02	30.08	0.28	636	0.76	0.72
Impératif	19.08	19.72	0.64	18.70	-0.38	4	0.24	0.00
Passé Simple	13.34	16.15	2.81	14.09	0.75	6	0.09	0.00
Plus-que-Parfait	21.27	23.44	2.17	23.22	1.95	17	0.51	0.25
Présent	27.55	27.97	0.42	27.59	0.04	2618	0.91	0.91
Subjonctif	26.81	27.72	0.91	26.07	-0.74	78	0.29	0.17
Passé Récent	24.54	30.53	5.96	30.08	5.54	3	0.22	0.00
Average/Total	23.31	25.21	1.89	24.30	0.99	3484	0.44	0.31

Table 7.16: BLEU and F1 scores per FR tense for baseline SMT, oracle SMT and SMT with tense predictions from the EXTENDED (Ext.) model. In bold are the values for the tenses where translation quality is improved the most. Compared to Table 7.15, the BLEU scores (and therefore translation quality) is higher due to better prediction performance on *Subjonctif* and *Imparfait*.

The results in terms of overall BLEU scores are shown in Table 7.14 for five predictor types. Then, in Tables 7.15 and 7.16, we show respectively the BLEU scores for two tense predictors combined with SMT, giving for each one the BLEU scores of the subsets of sentence sorted by expected tenses.

Labeling the verbs with the 9-CLASSES model in the test set prior to factored translation leads to an average improvement of +0.11 BLEU points overall. Moreover, the (unweighted) average improvement for the sentences actually containing a labeled verb is +0.33 BLEU points. Table 7.15 shows that the largest improvements can be obtained for infrequent tenses in French such as *Passé Simple* (+0.68 BLEU), *Plus-que-Parfait* (+0.80 BLEU), or *Passé Récent* (+2.02 BLEU). However, the rather poor labeling accuracy of the classifier for *Subjonctif* and *Impératif* leads to degraded translation quality of -0.70 and -1.01 BLEU, respectively. That was the reason why we tried to improve prediction with features specific to these tenses.

The scores of the factored model with model EXTENDED, in Table 7.16, show that this model has a higher labeling accuracy than model 9-CLASSES on *Imparfait* (+0.04 F1) and *Subjonctif* (+0.01

F1) tenses. Using this model therefore has a direct and measurable influence on verb tense translation quality (+0.12 BLEU points vs. the baseline) and leads to improvements on almost all tenses, compared to the one with 9-CLASSES: *Imparfait* (+0.16 BLEU), *Passé Composé* (+0.01 BLEU), *Impératif* (+0.63 BLEU, though still negative), *Passé Simple* (+0.07), *Plus-que-Parfait* (+0.85 BLEU), *Passé Récent* (+3.52 BLEU). Performance for *Présent* stays the same, while the *Subjonctif* is slightly degraded (-0.04 BLEU), likely due to minimal improvement of the labeling accuracy of only +0.01 F1 score. A better feature to capture when an EN verb should be translated to FR *Subjonctif* would be necessary but is not easy to find based on EN features only (as was mentioned in Chapter 6).

In order to check if the improvements for the BLEU scores as mentioned above are really due to verbs and their labeling we also performed manual evaluation of the translations output by the system that used the tense predictions of model EXTENDED. The evaluation criteria were the same as with the oracle experiment above (Section 7.1.2), and were applied to both the baseline and the tense-aware systems:

— Tense/Mode/Aspect (TAM): Are the TAM features correct, and if correct, are they the same or not as in the reference translation?
— Lexical choice: Are the lexical forms of the verbs output by the system correct, and if correct, are they the same as in the reference translation?
— Agreement (Person/Number): Is the verb-person-number agreement correct for the verbs output by the system?

	TAM			**Lexical choice**			**Agreement ok**		**Total VPs**
	Wrong	**Right**		**Wrong**	**Right**		**Yes**	**No**	
		≠ ref	**= ref**		**≠ ref**	**= ref**			
Baseline	206	61	387	47	267	340	536	118	654
	32%	9%	59%	7%	41%	51%	82%	18%	100%
Predicted	146	79	429	50	267	255	349	140	654
	22%	12%	66%	8%	39%	53%	79%	21%	100%

Table 7.17: Manual evaluation of a baseline and a tense-aware SMT system with labels from prediction model EXTENDED.

The scores of manual evaluation confirm what the BLEU scores revealed already: the tense-aware SMT system based on automated FR tense predictions performs at a level between the baseline and the oracle system, with high improvements over the baseline for tense/mode/aspect (+20%). Performance of lexical choice and agreement reaches the level of the oracle system described in Section 7.1.2, with scores given in Table 7.4.

7.9 Conclusions on factored translation models

The experiments presented in this chapter have shown that factored translation models are an effective and robust solution to incorporate discourse information into SMT. Still, there is a variety of ways to explore these methods in more detail in future work, as discussed also in the

conclusion of the thesis (Chapter 9).

To consider an even broader context than our classifiers and translation models do, labeling, for example, entire verbal phrase nodes in hierarchical or tree-based syntactical models could be considered. It will also likely prove useful to incorporate discourse features, for connectives and/or verb tenses, in document-wide decoding, where these features are directly modeled into new SMT decoding algorithms, as recently proposed by Hardmeier et al. [2012]. It has also been shown that it is difficult to choose the optimal parameters for factored translation models (Tamchyna and Bojar [2013]) and evaluating many configurations might lead to better results in translation quality.

In the following chapter, we explore a complementary way of dealing with discourse connectives in SMT. As we have already shown in the introductory chapter on translation problems (Chapter 2), human translators often have the choice of not translating connectives at all, or of inserting a target language connective where there was no source connective. To propose models for natural-sounding and coherent document-level SMT, the current techniques have to be extended in that direction as well, as we will do with the experiments illustrated in the following chapter.

8 Statistical machine translation with deletion/insertion of connectives

The first goal of this chapter is to analyze and measure semi-automatically, over larger corpora than manual analysis could cover, how many connectives are made implicit in human translation. Conversely, we also analyze which implicit relations are explicitated in human translation. These analyses are then compared to the translations produced by baseline SMT systems [1].

In these analyses of parallel corpora, we found that for implicitation, human translators tend to omit connectives more often than an SMT system does for the EN/FR and EN/DE language pairs. Even more often, human translators tend to insert a connective in the target language (FR and DE) for cases where there was no EN connective in the source text (Section 8.1).

The second goal is to show how discourse-aware and baseline SMT models can be tuned toward implicitating connectives (i.e. deleting them prior to translation) in a similar manner as human translators. Preliminary experiments show that such SMT systems can be obtained with a new tuning method in Moses: MIRA, which is based on sparse lexical features. In their best configurations, these SMT models increase the level of implicitation for connectives from 5% for the baseline system to up to 7-9% and therefore approach the human implicitation rate of 13%. The accuracy of implicitation, assessed manually, is higher for a system that implicitates connectives based on discourse relation labels assigned to them by the above-described classifiers (Section 8.2).

8.1 Semi-automatic corpus analyses for implicitation/explicitation of discourse connectives

The set and availability of discourse connectives varies in languages and as well does their (un-)ambiguity. It has been shown that connectives are difficult for language learners and in turn for translators, in terms of whether a language makes use more frequently of such markers or

1. This chapter contains work that has mostly been carried out during an internship of the author at the University of Edinburgh, supervised by Bonnie Webber, and that has been published in Meyer and Webber [2013].

not (e.g. see Spooren and Sanders [2008], Halverson [2004]). In language use and translation, one always needs to decide which and *if* a discourse marker should be used, depending on textual coherence structure, but also depending on style and genre of a text (Halverson [2004]). In the following, we will first focus on implicitation of discourse connectives in translation (Section 8.1.1), but the method used to detect this phenomenon is reversible and will be used to give examples of explicitation as well (Section 8.1.2).

8.1.1 Implicitation of connectives

Human translators can chose to *not* translate a source language discourse connective with a target language discourse connective, where the latter would be redundant or where the source language discourse relation would more naturally be conveyed in the target language by other means (cf. Figure 8.1). We will use the term 'zero-translation' or 'implicitation' for a valid translation that conveys the same sense as a lexically explicit source language connective, but not with the same form. As we will show, current SMT models either learn the explicit lexicalization of a source language connective to a target language connective, or treat the former as a random variation, realizing a connective word form or not. Learning other valid ways of conveying the same discourse relation might not only result in more fluent target language text, but also help raise the BLEU score by more closely resembling the more implicit human reference text.

EN: The man with the striking bald head was still needing a chauffeur, **1. as** the town was still unknown to him. **2. Otherwise** he could have driven himself — **3. after all**, no alcohol was involved and the 55-year-old was not drunk.
FR-REF: L'homme, dont le crâne chauve attirait l'attention, se laissa conduire **1. __0__** dans la ville qui lui était encore étrangère. **2. Autrement** notre quinquagénaire aurait pu prendre lui-même le volant — **3. __0__** il n'avait pas bu d'alcool et il n'était pas non plus ivre de bonheur.
DE-REF: Der Mann mit der markanten Glatze liess sich **1. wegen/Prep** der ihm noch fremden Stadt chauffieren. **2. Ansonsten** hätte er auch selbst fahren können — Alkohol war **3. schliesslich/Adv** nicht im Spiel, und besoffen vor Glück war der 55-jährige genauso wenig.

Figure 8.1: Examples of EN source connectives translated as zero or by other means in human reference translations.

Figure 8.1 is an excerpt from a news article in the newstest2010 data set (see Subsection on data below (p. 120). It contains two EN connectives — *as* and *otherwise* — that were annotated in the PDTB [2]. Using the set of discourse relations of the PDTB, *as* can be said to signal the discourse relation CAUSE (subtype 'Reason'), and *otherwise* the discourse relation ALTERNATIVE. This is discussed further in the subsection on the method (next page).

2. The excerpt contains a third possible connective *after all* that was not annotated in the PDTB, and our data as a whole contains other possible connectives not yet annotated there, including *given that* and *at the same time*. We did not analyse such possible connectives in what is described here..

The human reference translations do not translate the first connective *as* explicitly. In FR there is no direct equivalent, and the reason why the man needed a driver is given with a relative clause: *...dans la ville qui...* (lit.: in the town that was still foreign to him). In DE *as* is realized by means of a preposition, *wegen* (literally: because of). The second EN connective *otherwise*, maintains its form in translation to the target connective *autrement* in FR and *ansonsten* in DE.

On the other hand, baseline SMT systems for EN/FR and EN/DE (Section 8.1.1) both translated the two connectives *as* and *otherwise* explicitly by the usual target connectives, in FR: *comme, sinon* and in DE *wie, sonst*.

Method

The semi-automatic method that identifies zero- or non-connective translations in human references and machine translation output is based on a list of 48 EN discourse connectives with a frequency above 20 in the Penn Discourse TreeBank Version 2.0 (Prasad et al. [2008]). In order to identify which discourse relations are most frequently translated as zero, we have assigned each of the EN connectives the level-2 discourse relation that it is most frequently associated with in the PDTB corpus. The total list of EN connectives is given in Table A.2 in the Appendix of the thesis.

For every source connective, we queried its most frequent target connective translations from the online dictionary Linguee [3] and added them to dictionaries of possible FR and DE equivalents.

With these dictionaries and Giza++ word alignment (Och and Ney [2003]), the source language connectives can be located and the sentences of their translation (reference and/or automatic) can be scanned for an aligned occurrence of the target language dictionary entries. If more than one discourse connective appears in the source sentence and/or a discourse connective is not aligned with a connective or connective-equivalent found in the dictionaries, the word position (word index) of the source language connective is compared to the word indexes of the translation in order to detect whether a target language connective (or connective-equivalent from the dictionaries) appears in a 5-word window to its left and right (the method extends on the ACT metric, Chapter 4). This also helps filtering out cases of non-connective uses of e.g. *separately* or *once* as adverbs. Finally, if no aligned entry is present and the alignment information remains empty, the method counts a zero-translation and collects statistics on these occurrences.

After a first run where we only allowed for actual connectives as translation dictionary entries, we manually looked through 400 cases for each, FR and DE reference translations, that were output as zero-translations (in the newtest2012 data, p. 120). We found up to 100 additional cases that actually were not implicitations, but conveyed the source language connective's

3. http://www.linguee.com

meaning by means of a paraphrase, e.g. EN: *if* – FR: *dans le cas où* (lit.: in case where) – DE: *im Falle von* (lit.: in case of). For example, the EN connective *otherwise* ended up with the dictionary entries in Figure 8.2.

EN: **otherwise** ALTERNATIVE :
FR: autrement|sinon|car|dans un autre cas|d'une autre manière
DE: ansonsten|andernfalls|anderenfalls
|anderweitig|widrigenfalls|andrerseits| andererseits|anders|sonst

Figure 8.2: Dictionary entries of FR and DE connectives and equivalents for the EN connective *otherwise*.

Data

For the experiments described here, we concatenated two data sets, the newstest2010 and newstest2012 parallel texts as publicly available by the Workshop on Machine Translation [4]. The texts consist of complete articles from various daily news papers that have been translated from EN to FR, DE and other languages by translation agencies.

In total, there are 5,492 sentences and 117,799 words in the source language texts, of which 2,906 are tokens of the 48 EN connectives. See Table A.2 for the connectives and their majority class, which aggregate to the detailed statistics given in Table 8.1.

Rel.	**TC**	**Rel.**	**TC**
Alternative	30	Conjunction	329
Asynchrony	588	Contrast	614
Cause	308	Instantiation	43
Concession	140	Restatement	14
Condition	159	Synchrony	681

Table 8.1: Total counts (TC) of English discourse connectives (2,906 tokens) from the newstest2010+2012 corpora, whose majority sense conveys one of the 10 PDTB level-2 discourse relations (Rel.) listed here.

To produce machine translations of the same data sets we built EN/FR and EN/DE baseline phrase-based SMT systems, by using the Moses decoder (Koehn et al. [2007]), with the Europarl corpus v7 (Koehn [2005]) as training and newtest2011 as tuning data. The 3-gram language model was built with IRSTLM (Federico et al. [2008]) over Europarl and the rest of WMT's news data for FR and DE.

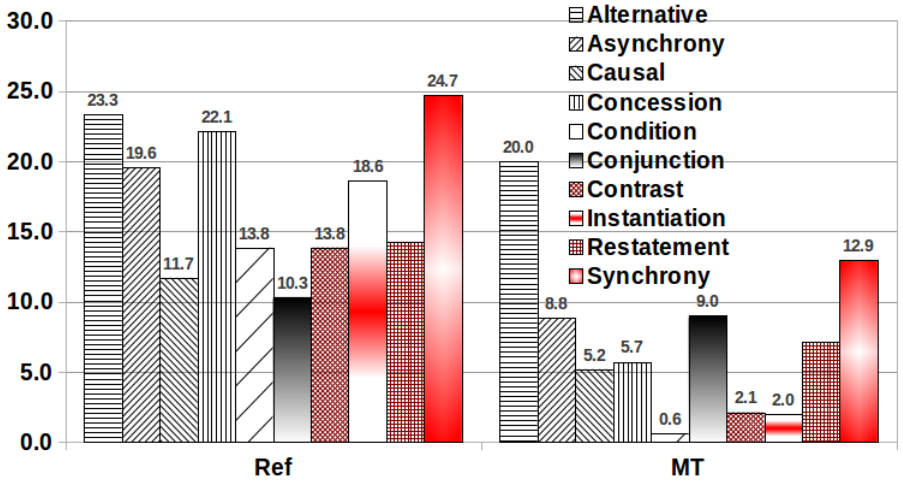


Figure 8.3: Percentage of zero-translations in newstest2010+2012 for EN/FR per discourse relation and translation type: human reference (Ref) or MT output (MT).

Results

In order to group the individual counts of zero-translations per discourse connective according to the discourse relation they signal, we calculated the relative frequency of zero-translations per relation as percentages, see Figures 8.3 for EN/FR, and 8.4 for EN/DE. The total percentage of zero-translations in the references and the baseline MT output is given in Table 8.2.

A first observation is that an MT system seems to produce zero-translations for discourse connectives significantly less often than human translators do. Human FR translations seem to have a higher tendency toward omitting connectives than the ones in DE. Figures 8.3 and 8.4 also show that the discourse relations that are most often rendered as zero are dependent on the target language. In the FR reference translations, SYNCHRONY, ALTERNATIVE and CONCESSION account for most implicitations, while in the DE reference translations, CONDITION, ALTERNATIVE and CONCESSION are most often left implicit.

Translation	Type	C	%
EN/FR	Ref	508	17.5
	MT	217	7.5
EN/DE	Ref	392	13.5
	MT	129	4.4

Table 8.2: Counts (C) and relative frequency (%) of zero-translations for EN/FR and EN/DE in human references (Ref) and MT output (MT) over newstest2010+2012.

The results are to some extent counterintuitive as one would expect that semantically dense

4. http://www.statmt.org/wmt12/

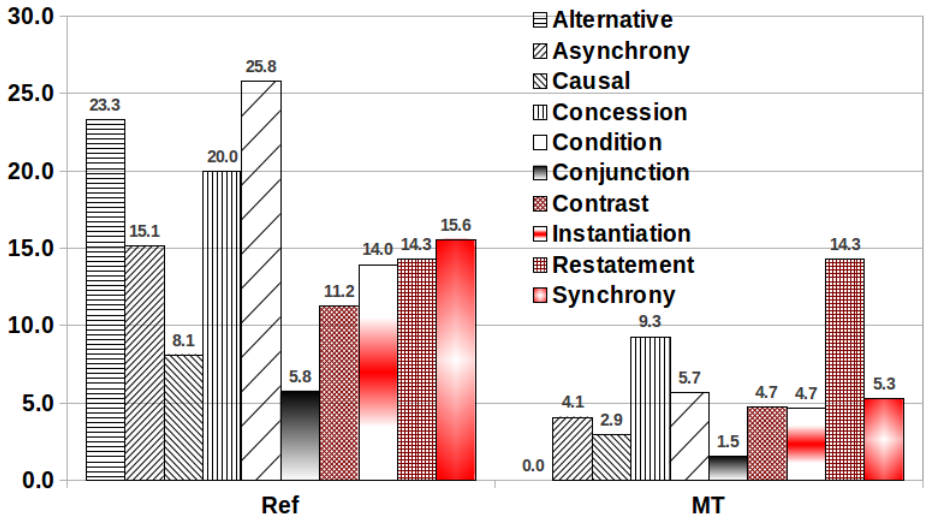


Figure 8.4: Percentage of zero-translations in newstest2010+2012 for EN/DE per discourse relation and translation type: human reference (Ref) or MT output (MT).

discourse relations like CONCESSION would need to be explicit in translation in order to convey the same meaning. A subsection (p. 123) presents some non-connective means available in the two target languages, by which the discourse relations are still established.

We furthermore looked at the largest implicitation differences per discourse relation in the human reference translations and the MT output. For EN/FR for example, 13.8% of all CONDITION relations are implicitated in the references, by making use of paraphrases such as *dans le moment où* (lit.: in the moment where) or *dans votre cas* (lit.: in your case) in place of the EN connective *if*. The MT system translates *if* in 99.4% of all cases to the explicit FR connective *si*. Similarly, for INSTANTIATION relations and the EN connective *for instance* in the references, the translators made constrained use of verbal paraphrases such as *on y trouve* (lit.: among which we find). MT on the other hand outputs the explicit FR connective *par exemple* in all cases of *for instance*.

For EN/DE, there is the extreme case, where ALTERNATIVE relations are, in human reference translations, quite often implicitated (in 23.3% of all cases), whereas the MT system translates all the instances explicitly to DE connectives: *wenn* (unless), *sonst* (otherwise) and *statt, stattdessen, anstatt* (instead). The translators however make use of constructions with a sentence-initial verb in conditional mood (cf. Section 8.1.1) for *otherwise* and *unless*, but not for *instead*, which is, as with MT, always explicitly translated by humans, most often to the DE connective *statt*. The very opposite takes place for the RESTATEMENT relation and the EN connective *in fact*. Here, MT leaves implicit just as many instances as human translators do, i.e. 14.3% of all cases. Translators use paraphrases such as *in Wahrheit* (lit.: in truth) or *übrigens* (lit.: by the way), while the translation model tends to use *im Gegenteil* (lit.: opposite), which is not a literal translation of *in fact* (usually *in der Tat* or *tatsächlich* in DE), but reflects the

contrastive function this marker frequently had in the Europarl training data of the baseline MT system.

Case studies

Temporal connectives from EN to FR The most frequent implicitated discourse relation for EN/FR translation is SYNCHRONY, i.e. connectives conveying that their arguments describe events that take place at the same time. However, since the situations in which SYNCHRONY relations are implicitated are similar to those in which CONTRAST relations are implicitated, we discuss the two together.

We exemplify here cases where EN discourse connectives that signal SYNCHRONY and/or CONTRAST are translated to FR with a '*en*/Preposition + Verb in Gerund' construction without a target language connective. The EN source instances giving rise to such implicitations in FR are usually of the form 'discourse connective + Verb in Present Continuous' or 'discourse connective + Verb in Simple Past', see sentences 1 and 2 in Figure 8.5.

1. EN: In her view, the filmmaker "is asking a favour from the court, **while** at the same time **showing** disregard for its authority".
FR-REF: Pour elle, le cinéaste "demande une faveur à la cour, tout **en/Prep méprisant/V/Ger** son autorité".
FR-MT*: Dans son avis, le réalisateur de "demande une faveur de la cour, **alors que** dans le même temps une marque de mépris pour son autorité".
2. EN: **When** Meder **looked** through the weather-beaten windows of the red, white and yellow Art Nouveau building, she could see weeds growing up through the tiles.
FR-REF: **En/Prep jetant/V/Ger** un coup d'œil par la fenêtre de l'immeuble-art nouveau en rouge-blanc-jaune, elle a observé l'épanouissement des mauvaises herbes entre les carreaux.
FR-MT*: **Lorsque** Meder semblait weather-beaten à travers les fenêtres du rouge, jaune et blanc de l'art nouveau bâtiment, elle pourrait voir les mauvaises herbes qui grandissent par les tuiles.

Figure 8.5: Translation examples for the EN temporal connectives *while* and *when*, rendered in the FR reference as a 'preposition + Verb in Gerund' construction. MT generates the direct lexical equivalents *alors que* and *lorsque*.

Out of 13 cases of implicitations for *while* in the data, 8 (61.5%) have been translated to the mentioned construction in FR, as illustrated in the first example in Figure 8.5, with a reference and machine translation from newstest2010. The discourse connective *while* here ambiguously signals SYNCHRONY and/or CONTRAST, but there is a second temporal marker (*at the same time*, a connective-equivalent not yet considered here or in the PDTB), that disambiguates *while* to its CONTRAST sense only or to the composite sense SYNCHRONY/CONTRAST. The latter is conveyed in FR by *en méprisant*, with CONTRAST being reinforced by *tout* (lit.: all).

In Example 2, from newstest2012, the sentence-initial connective *when*, again signaling SYNCHRONY, is translated to the very same construction of '*en*/Preposition + Verb in Gerund' in the FR reference.

In the baseline MT output for Example 1, neither of the two EN connectives is deleted, *while* is literally translated to *alors que* and *at the same time* to *dans le même temps*. While the MT output is not totally wrong, it sounds disfluent, as *dans le même temps* after *alors que* is not necessary.

In the baseline MT output for Example 2, the direct lexical equivalent for *when* – *lorsque* is generated, which is correct, although the translation has other mistakes such as the wrong verb *semblait* and the untranslated *weather-beaten*.

To model such cases for SMT one could use POS tags to detect the 'discourse connective + Present Continuous/Simple Past' in EN and apply a rule to translate it to 'Preposition + Gerund' in FR. Furthermore, when two connectives follow each other in EN, and both can signal the *same* discourse relations, a word-deletion feature (as it is available in the Moses decoder via sparse features), could be used to trigger the deletion of one of the EN connectives, so that only one is translated to the target language (see Section 8.2). Another possibility would be to treat cases like *while at the same time* as a multi-word phrase that is then translated to the corresponding prepositional construction in FR.

Conditional connectives from EN to DE Out of the 41 cases involving a CONDITION relation (10.5% of all DE implicitations), 40 or 97.6% were due to the EN connective *if* not being translated to its DE equivalents *wenn*, *falls*, *ob*. Instead, in 21 cases (52.5%), the human reference translations made use of a verbal construction which obviates the need for a connective in DE when the verb in the *if*-clause is moved to sentence-initial position and its mood is made conditional, as in Figure 8.6, a reference translation from newstest2012, with the DE verb *wäre* (lit.: were) (VMFIN=modal finite verb, Konj=conditional). This construction is also available in EN (*Were you here, I would...*), but seems to be much more formal and less frequent than in DE where it is ordinarily used across registers. In the baseline MT output for this sentence, *if* was translated explicitly to the DE connective *wenn*, which is in principle correct, but the syntax of the translation is wrong, mainly due to the position of the verb *tun*, which should be at the end of the sentence.

The remaining 19 cases of EN *if* were either translated to DE prepositions (e.g. *bei, wo*, lit.: at, where) or the CONDITION relation is not expressed at all and verbs in indicative mood make the use of a conditional DE connective superfluous.

Of the 21 tokens of *if* whose reference translations used a verbal construction in DE, 14 (66.7%) were tokens of *if* whose argument clause explicitly referred to the preceding context – e.g., *if they were, if so, if this is true* etc. These occurrences could therefore be identified in EN and could be modeled for SMT as re-ordering rules on the verbal phrase in the DE syntax tree after

EN: **If** not for computer science, they would be doing amazing things in other fields.
DE-REF: **__0__ Wäre/VMFIN/Konj** es nicht die Computerbranche gewesen, würden sie in anderen Bereichen fantastische Dinge schaffen.
DE-MT*: **Wenn** nicht für die Informatik, würden sie tun, erstaunlich, Dinge auf anderen Gebieten.

Figure 8.6: Translation example for the EN connective *if*, rendered in the DE reference as a construction with a sentence-initial verb in conditional mood. MT generates the direct lexical equivalent *wenn*.

constituent parsing in syntax-based translation models.

8.1.2 Explicitation of connectives

Method and data

The algorithm of word alignment and its dictionary refinement (described in Section 8.1.1) as well as the data sets (newstest2010+2012) used in the analysis of implicitation can be kept for detecting cases of target language connectives for which there was no equivalent in the source language.

What does need to change are the dictionaries. When the goal is to detect explicitation of connectives in EN/FR and EN/DE translation (i.e. EN remains the source), we have to build two new dictionaries, one that contains French connectives, the relations they signal along with valid EN translations and paraphrases and the same holds true for a dictionary of DE connectives. Based on these 'inverted' dictionaries we can then find target language connectives in FR and DE and check whether they have been aligned to an EN connective or a valid paraphrase and if not, count a case of explicitation.

Instead of the PDTB, we made use of two resources for connectives in FR (LexConn, Roze et al. [2010]) and DE (DimLex, Stede and Umbach [1998]), respectively. Both resources provide an XML-formatted lexicon of about 300 connectives, the senses they can signal and examples of their usage.

In FR, we took into account 105 of such markers and there are 2744 occurrences of these in newstest2010+2012. For DE we considered 95 discourse connectives with 3816 instances in newstest2010+2012. The detailed statistics of the discourse relation distribution is given in Table 8.3 for EN/FR and 8.4 for EN/DE, respectively.

With LexConn and DimLex, there are no frequency indications, neither for the connectives, nor for the discourse relations because they are mere lexicons as opposed to the PDTB which provides a fully annotated corpus. We could therefore not focus on smaller sets of the most frequent connectives in FR and DE. It should also be noted, that these two resources might have a broader definition of discourse connectives than the PDTB which is why more markers are

Rel.	TC	Rel.	TC
Alternation	37	Explanation	508
Background	81	Flashback	99
Background-Inverse	142	Flashback-Explanation	8
Concession	93	Goal	72
Condition	227	Narration	157
Consequence	5	Narration-Result	150
Continuation	143	Parallel	20
Contrast	529	Result	210
Detachment	9	Evidence	8
Elaboration	54	Violation	193

Table 8.3: Total counts (TC) of French discourse connectives (2,744 tokens) from the newstest2010+2012 corpora, with assigned discourse relations from the LexConn resource for French connectives.

Rel.	TC	Rel.	TC
Asymmetric-Contrast	3	Elaboration	637
Cause	401	Joint	216
Circumstance	642	Means	22
Concession	238	Not-Yet	440
Condition	24	Pre-Condition	56
Contrast	465	Sequence	672

Table 8.4: Total counts (TC) of German discourse connectives (3'816 tokens) from the newstest2010+2012 corpora, with assigned discourse relations from the DimLex resource for German connectives.

considered and these may already include what we above called paraphrases, i.e. multiword expressions such as *at the same time – en même temps – zur selben Zeit*. The sense inventories in LexConn and DimLex are inspired by the Rhetorical Structure Theory (Mann and Thompson [1988]) and therefore somewhat comparable to the PDTB, but RST relations are more fine-grained and numerous. In FR, we considered 20 RST relations, in DE 12. Due to the lack of frequency indications for the relations, we assigned, to each connective, the discourse relation as given in the lexicons, if several relations are possible, we decided ourselves upon the likely most frequent one.

The EN translations were taken from inverting the dictionaries for implicitation (Section 8.1.1), to enrich them we made again use of www.linguee.com and manually went through several hundred cases in order to complete the dictionaries with possible paraphrases. The list of connectives and relations considered is given in Table A.3 (FR) and A.4 (DE) in the appendices of the thesis, respectively. An FR and DE dictionary example is given in Figures 8.7 and 8.8.

FR: malgré tout VIOLATION
EN: despite|in spite of|albeit|after all|notwithstanding|nonetheless|nevertheless

Figure 8.7: Dictionary entry for the FR connective *malgré tout* and its EN translations.

DE: vielmehr ASYMMETRIC CONTRAST
EN: but|rather|in reality|in point of fact|in truth|in fact|on the contrary

Figure 8.8: Dictionary entry for the DE connective *vielmehr* and its EN translations.

Results

Unlike with the implicitation results presented above (Section 8.1.1), we here cannot draw direct comparisons of EN/FR and EN/DE translations, as the set of connectives and discourse relations is not based on EN as source language anymore. In the two target languages we separately detected which connectives have been inserted without having had an equivalent EN source language expression.

What however is comparable, is the amount by which a baseline SMT system does not explicitate discourse connectives as opposed to human reference translations: For EN/FR this is about 3 times less (26.46% of explicitation in human reference translations vs. 8.01% in SMT output), while for EN/DE, explicitation in SMT happens even less frequently, i.e. about 7 times less (30.84% vs. 4.38%), as can be seen in Table 8.5.

Translation	Type	C	%
EN/FR	Ref	726	26.46
	MT	221	8.01
EN/DE	Ref	1177	30.84
	MT	167	4.38

Table 8.5: Counts (C) and relative frequency (%) of explicitated discourse connectives in EN/FR and EN/DE translations by humans (Ref) and MT output (MT) over newstest2010+2012.

We again analyzed explicitation rates per discourse relation and summarize the results in Figures 8.9 for EN/FR and 8.10 for EN/DE, respectively. For EN/FR the first figure shows that the discourse relations of GOAL, CONTINUATION and RESULT are explicitated the most. As figure 8.10 for EN/DE illustrates, the three most often explicitated relations are ASYMMETRIC-CONTRAST, ELABORATION and CAUSE, which are of clearly different semantic classes than those for EN/FR. As with implicitation, the explicitation of connectives is dependent on the actual translation direction. In the next subsection, we will analyze concrete translation examples with the most frequently explicitated discourse relations in both language pairs.

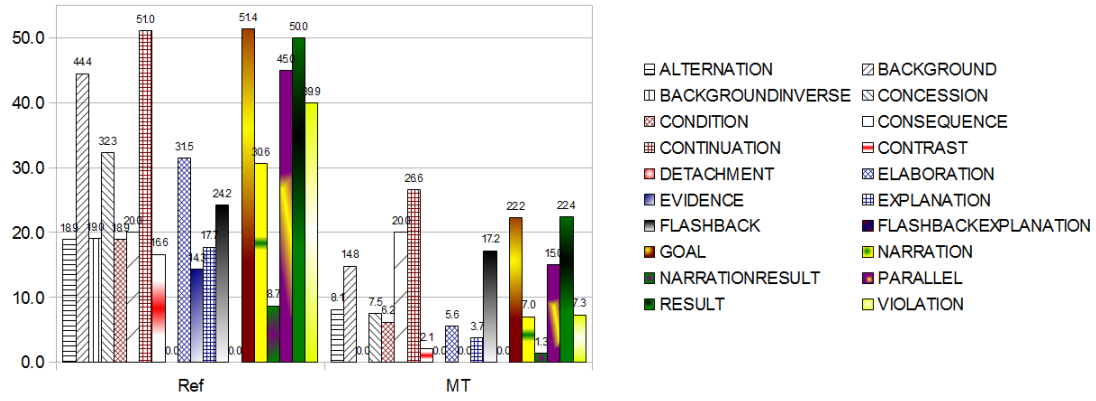


Figure 8.9: Percentages of explicitation of connectives in newstest2010+2012 for EN/FR per discourse relation and translation type: human reference (Ref) or MT output (MT).

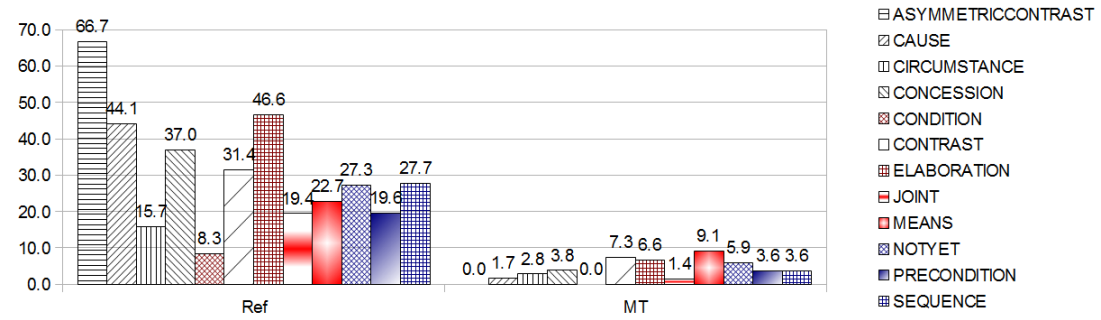


Figure 8.10: Percentages of explicitation of connectives in newstest2010+2012 for EN/DE per discourse relation and translation type: human reference (Ref) or MT output (MT).

Examples

As Figure 8.9 shows, in human EN/FR reference translations, connectives that can signal GOAL in French are most often inserted where there was no connective in English. In 95% of all cases of explicitation with GOAL, this is due to the FR connectives *pour que* and *afin de*, which are frequently triggered by the EN preposition *to* (not a connective in EN). More literal equivalents in EN for *pour que* and *afin de* would be the connectives so that or the connective-like paraphrase *in order to*. Vice-versa, in FR, one could use only *pour* in order to more literally translate the EN preposition *to*. The EN/FR baseline system seems to have learned the frequent correspondence of *to* – *pour que*|*afin de* quite reasonably, as we found that 43.75% of the *pour que* and *afin de* occurrences generated by MT correspond to cases where there has been the preposition *to* in EN.

In Figure 8.11 we illustrate a more problematic case for the FR connective *en outre*, signaling the second most frequently explicitated discourse relation of CONTINUATION. As there is no

surface word or paraphrase that would signal this relation in the EN source, the SMT system cannot generate such a connective, whereas the human translator used it at the beginning of the sentences. In plus, even the sentence-initial *comme* in the FR reference can be regarded as discourse marker (signaling EXPLANATION) that has no equivalent in EN and is again not translated at all in the FR MT output. When scoring with automatic tools such as BLEU, in the FR SMT output, bi- to tri-grams are missing when compared to the reference translation which significantly affects the score.

EN: __**0**__ The police announced in a statement that one 19-year-old has been arrested subsequently on suspicion.
FR-REF: Comme **en outre** informé par la police, un peu plus tard, un homme de 19 ans a été pu être arrêté en vertu de soupçons.
FR-MT: __**0**__ La police a annoncé dans une déclaration que l' un de 19 ans a été arrêté par la suite sur la base de soupçons.

Figure 8.11: Example for explicitation by a human translator with the FR connective *en outre* that has no equivalent in the EN source text and is therefore omitted in the SMT output.

In the following, we similarly look at a case for EN/DE translation with an explicitated ASYMMETRIC-CONTRAST relation, signaled by the DE connective *vielmehr* that can literally be translated in EN by *rather* or *in fact*. In Example 8.12, the human DE reference translation makes use of *vielmehr* to explicitate the contrast that the person described is merely professor than banker. In the EN source, the only particle pointing to that contrast is the negating *not* and consequently, the SMT system only translates this as *nicht*.

EN: He is an economics professor and central banker , **not** a conventional banker , and clearly would need some time to adjust.
DE-REF: Er ist kein Banker, **vielmehr** Oekonomieprofessor und Notenbanker und haette eine Einarbeitungszeit sicherlich noetig gehabt.
DE-MT Er ist ein Wirtschaft Professor und Zentralbanker, **nicht** auf ein konventionelles Banker, und einige Zeit brauchen wuerde, sich anzupassen.

Figure 8.12: Example for explicitation by a human translator with the DE connective *vielmehr* that has no equivalent in the EN source text and is therefore omitted in the SMT output.

Other frequently explicitated discourse connectives in DE are *so* (EN: *so*) and *nämlich* (EN: *namely, it is*), both signaling ELABORATION. They therefore are more like 'fillers' that are naturally used in DE to continue or stress ongoing explanations and descriptions. Usually there is no equivalent at all in EN, in our data for *nämlich* there is no marker in EN in the total of 22 cases found, and for *so*, 27% of all cases were implicit in EN, but explicit in DE. Again, if an SMT system would learn to insert these in the correct places, automatic scoring methods would find more n-grams to match between system output and human reference and in turn, the system output would score higher and become more fluent.

In the following section we describe SMT experiments with features that trigger source word deletion (implicitation) and specifically tune these models toward omission of discourse connectives where appropriate.

8.2 Sparse lexical features for SMT

In order to model, for machine translation, the 'natural' deletion and insertion of connectives as they are performed by human translators, one needs to find, possibly without inserting any hand-crafted rules, a way to force an SMT system to either generate a target word or paraphrase where there was no source language equivalent and/or to suppress a source language connective prior to decoding.

One approach to this would be to train a classifier, similar to the ones used in this thesis, but instead of a predicted discourse relation as output, there should be a likelihood value of how probable it is whether a discourse relation should be expressed by the means of a lexically explicit discourse connective or whether this relation can be inferred from the context implicitly. Such a model has only been publishd recently: Patterson and Kehler [2013] build a logistic regression classifier, that, based on features annotated in the PDTB, predicts whether a discourse relation between two clauses is more likely to be realized by an explicit lexical connective, or whether the connective is to be omitted and the relation is present implicitly. Although the model has a high accuracy to predict this likelihood (almost 87%), the 'disadvantage' is that it relies heavily on argument and other context features that are part of the PDTB monolingual annotation only and are therefore not available immediately in other texts or corpora, and especially not in the parallel texts that are used to train SMT systems. We therefore tried to approach the problem directly in the translation models.

8.2.1 SMT tuning with lexical features

With newest developments for MT tuning algorithms it has become possible to integrate a multitude of translation features, i.e. basically one translation feature for every single word. These features can be used to decide whether a word should be deleted from and/or inserted in a text. MIRA stands for Margin Infused Relaxed Algorithm (cf. Watanabe et al. [2007]) and is, as opposed to MERT (see Chapter 3), an online learning method which allows to include many more features (i.e. millions), as only the actually active ones need to be updated (online) instead of precomputed offline as with standard MERT (which is why MERT can only reliably and scalably be used for about a dozen of features). As with MERT, the loss function to be optimized can be the BLEU score, but MIRA measures the difference between a correct (so-called 'hope') and incorrect (so-called 'fear') translation according to the reference translation(s). A larger error (or margin) then means a larger distance between the scores of the correct and incorrect translations.

Due to online learning, MIRA allows for sparse feature coding, i.e. for each source or target

language word, a single feature can be (de-)activated and in turn, based on its weight, the source or target language word can be deleted and/or inserted prior to or after decoding respectively. In the Moses SMT toolkit (Koehn et al. [2007]), three types of such sparse lexical features have been considered at the time of writing [5]:

— `wt`: word translation, a feature which indicates if a specific source word should be translated by a specific target word
— `twi`: target word insertion, i.e. a specific target word has no alignment point and does not align to a source word in the alignment stored with the translation model
— `swd`: source word deletion, indicates whether a specific source word has an alignment point or not

While it is quite obvious that this is useful for example to delete an EN word such as 'the' when translating to a language without or infrequent grammatical articles, it is probably not useful to delete many content words such as nouns as the contained information has most often to be conveyed to the target language. To the best of our knowledge, we are the first applying these new tuning methods to discourse connectives, which are function words, but still carry and signal semantic information.

For the SMT experiments we here focus on EN/DE translation and on cases where an explicit source connective should be implicitated (i.e. deleted) so that the target translation does not contain a connective anymore while still being coherent (cf. Section 8.1.1). We experimented with different configurations of translation models with `swd` features in order to see whether the weights that are learned for omission regarding connectives would lead to more correct or accurate translations.

In the following we discuss the experimental setting, the models built and the translation results obtained.

8.2.2 Data

We will compare the `swd` models against the same human reference translations, baseline and factored SMT systems for EN/DE SMT as described in Section 7.7.2, except that we no longer consider the nt2008+sy2009 test set (in order to compare directly to the implicitation rates of the reference translations that were computed over nt2010+nt2012 (Section 8.1.1). The data is given again in Table 8.6.

The baseline system was built without any modification to the text from the corpora, except for tokenization and true-casing with the Moses tools. The language model was a 3-gram one built with IRSTLM (Federico et al. [2008]) over a combination of Europarl v7 and the News Commentary corpus (years 2007-2011), as distributed by the workshops on statistical MT.

5. The term 'sparse' here accounts for the fact that considering a specific translation feature for each word is much rarer to be activated compared to the overall translation and language model probability features.

Language pair	Role	Data source	Genre	# Sentences	# Labeled connectives
EN/DE	training	EP	parliamentary debates	1,906,486	133,448
	tuning (1)	nt2011	newswire	3,003	174
	testing (2)	nt2012	newswire	3,003	176
	testing (3)	nt2010	newswire	2,489	165

Table 8.6: Genres, sizes and numbers of connectives in the data for training, tuning and testing SMT systems. The sources are: EP (Europarl corpus v. 7), nt (newstest).

For the model using labels, the labeling of the 7 EN discourse connectives (*although, however, meanwhile, since, (even) though, while, yet*) in training/tuning/testing data was done by using the MaxEnt classifier and the `All_features` model (Table 5.10, Section 5.6).

8.2.3 Models

Based on the EN/DE data described above, we then built a series of SMT models that make use of sparse lexical features (`swd` features), that consider certain words to be deleted during decoding. Practically, one has to provide the decoder with word lists, that contain source language words that are likely candidates for deletion. We show the system configuration and and the discourse connectives used in each configuration in Table 8.7. An example source word deletion list (with already tuned weights as output by MIRA) is given in Figure 8.13. The 7 connectives are the same as have been considered in our classifiers: *although, however, meanwhile, since, (even) though, while, yet*. When context words are included, we extracted 3 words preceding and following the connective, sorted them by frequency and limited the set to 500 words. Context words surrounding the connectives are likely candidates for implicitation as well, as they might consist of other connectives and connective-like paraphrases, e.g. *. . . **but** since . . . , . . . while **at the same time***. For configuration 5 in Table 8.7, the 48 connectives are the same that were used to detect implicitation in Table A.2. For these there is no model including the 500 context words, as the 48 connectives here already contain potential multi-word expressions.

All tunings with sparse features (MIRA) and without them (the baseline system is tuned with MERT to compare against) were repeated 3 times to gain stable scores (both tuning methods inhibit randomness).

8.2.4 Results and discussion

In order to evaluate the above-described EN/DE models, we computed the BLEU score of all models, by averaging over the three tuning runs. A second score is the 'implicitation rate' for each model (from the method described in Section 8.1.1 that computes the statistics on zero-translations in the outputs of each model). Figure 8.14 summarizes the BLEU scores and implicitation rates for the human reference translations, the baseline translations and the `swd` models.

Model	Configuration	Tuning	`swd` features
1	Human ref.	–	–
2	Baseline SMT	MERT	–
3	`swd-1`	MIRA	7 conn., unlabeled
4	`swd-2`	MIRA	7 conn., 500 surrounding words
5	`swd-3`	MIRA	48 conn.
6	`swd-4`	MIRA	7 *labeled* conn.
7	`swd-5`	MIRA	7 *labeled* conn., 500 surrounding words

Table 8.7: System configuration for SMT models with source word deletion features (`swd`) in order to compare against the implicitation of connectives in human reference translations and a baseline SMT system that was tuned without the deletion features.

swd_while	0.00363853973998774
swd_little	1.70250410661525e-06
swd_its	0.0110404619795686
swd_account	8.05114757634332e-06
swd_for	-0.025059502971019
swd_policies	-1.47365588019115e-05
swd_not	0.00979219275599005
swd_the	0.00174179179274332
swd_clearly	0.000153514964978291
swd_today	0.000322289117926106
swd_areas	-0.00109862759684677
swd_production	-1.00932362063903e-07
swd_responsibility	0.000176630558941465
swd_appropriate	4.66661466167079e-05
swd_especially	0.000159056870372516
swd_is	-0.0215049458920778
swd_other	0.021779269643476
swd_you	0.000597416044019059
swd_forward	0.000465720281172592
swd_will	0.0225222627977782
swd_mentioned	-4.95803132805296e-06
swd_known	-0.00094385017699606
swd_billion	-0.00114657146996741
swd_'	-0.000138379081292901
swd_fact	4.25666751004535e-05
swd_cooperation	0.000659585159093647

Figure 8.13: Example excerpt of a source word deletion list for the connective *while* and surrounding words, to which feature weights (deletion probabilities) are assigned by the MIRA tuning algorithm. The lower these weights, the less likely it is that word should be deleted during translation.

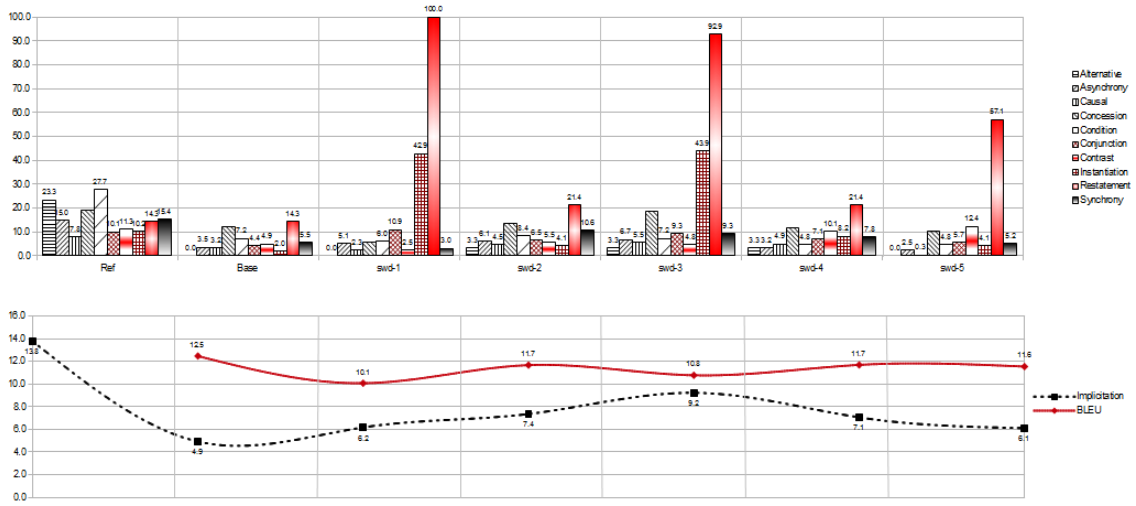


Figure 8.14: Implicitation of discourse connectives in human reference translations, and translations output by a baseline SMT system (base) compared to SMT models with sparse lexcial source word deletion features, under different configurations. While keeping the BLEU score more or less stable, the curves show that the implicitation rate for the `swd` models is higher than for the baseline system and approaches the reference percentage.

The graph shows that the `swd` models *increase* the level of implicitation and indeed approach the human level of omitting connectives in translation. The highest implicitation rate (9.2% vs. the 13% of the human reference) is, quite obviously, reached in configuration `swd-3`, when considering the most connectives (48 EN ones), although the BLEU score is degraded the most compared to the baseline SMT system (10.8 vs. 12.5). The graph further shows that the model using 7 **labeled** connectives only (`swd-4`) can reach a high implicitation performance (7.1%) and has a higher BLEU score of 11.6. We have therefore compared `swd-3` and `swd-4` to each other as well as each of them against the baseline and the human reference, by manually evaluating the actual zero-translations output by the models (the first 50 translations of model `swd-3` and the first 50 for model `swd-4`).

We applied a similar evaluation strategy as with measuring $\Delta Connectives$ (see the previous chapter, 7), i.e. we compare the translations output by the modified models (`swd-3` and `swd-4`) against the baseline system, not configured for implicitation, nor using any labeled connectives. Thereby, we judge whether the modified system output is better (+), equal (=) or worse (−) than the baseline. The only difference here is that there likely is no actual, lexically explicit connective anymore in the `swd-3` and `swd-4` translations, as this was exactly the goal: to implicitate the connective and to see whether a coherent, acceptable translation would result. We therefore do not consider connective translation quality but evaluated for overall translation quality ($\Delta Readability$ below). Figure 8.15 illustrates the three cases: 1. a translation by model `swd-4` with an implicitated connective and better readability than its baseline counterpart; 2. a translation by model `swd-3` with equal readability and 3. a

translation that has been degraded by the swd-4 model when compared to the baseline.

1. EN SOURCE: But **since then**, there has been almost no tolerance of criticism by the authorities, our correspondent says.
SWD-4, DE TRANS.: Aber **seitdem __0__**, hat es fast keine Toleranz der Kritik von die Behörden, unsere Korrespondenten sagt.
BASE, DE TRANS.: Aber ***da dann**, es hat worden fast keine Toleranz der Kritik von den Behörden , unsere Korrespondent sagt.
DE REF: Aber **seitdem** gab es seitens der Behörden keinerlei Toleranz mehr für Kritik, sagte unser Korrespondent.

2. EN SOURCE: It's not just that in the text itself he repeatedly uses the words 'just like in Iraq', but he even went **as far as** to use the name of Bush's declaration from January 2007 as the title of his own declaration: 'the new way forward'.
SWD-3, DE TRANS.: Es nicht nur, dass in der Text selbst er wiederholt die Worte 'nur wie in Irak', aber er ging sogar **so weit**, die Namen der Bush Erklärung von Januar 2007 als der Titel seiner eigenen Erklärung 'die neuen Weg'.
BASE, DE TRANS.: Es des nicht nur, dass in der Text selbst er wiederholt nutzt, die Worte 'nur wie im Irak', aber er gingen sogar **so weit**, wie zu nutzen die Namen der Bush Erklärung des aus Januar 2007 wie der Titel seiner eigenen Erklärung: 'die neuen Weg'.
DE REF: Nicht nur, dass er in dem Text selbst mehrmals die Redensart 'so wie im Irak' verwendet, er hat sogar **__0__** den Namen der Bush-Erklärung vom Januar 2007 als den Namen seiner eigenen Deklaration: 'neuer Vorwärtsweg' verwendet, ohne zu zaudern.

3. EN SOURCE: Czech railways have concluded a new ten-year contract for local and express trains, **whereas**, **previously**, the contract was always for one year.
SWD-4, DE TRANS.: Tschechischen Eisenbahn abgeschlossen haben eine neue zehnjährige Vertrag für lokalen und zum Ausdruck bringen Züge, **während**, ***__0__** der Vertrag war immer für ein Jahr.
BASE, DE TRANS.: Tschechischen Eisenbahn haben abgeschlossen eine neue von zehn Jahren Vertrag für lokalen und zum Ausdruck bringen Züge, **während**, **zuvor**, den Vertrag war immer für ein Jahr.
DE REF: Für die Lokalzüge sowie die Schnellzüge hat die Eisenbahn einen Zehnjahresvertrag neu abgeschlossen, **__0__ bis jetzt** wurde der Vertrag immer nur für ein Jahr abgeschlossen.

Figure 8.15: Examples of implicitated connective translations with models swd-4 (1., 3.) and swd-3 (2.) that are better (1.), equal (2.) or worse (3.) than their baseline counterpart.

Example 1 : In example sentence 1. there are two explicit EN connectives: *since* and *then*. The second, *then*, is, when translating to German, rather superfluous and can already be expressed with the translation for *since*, that here signals a TEMPORAL discourse relation:

seitdem, as in the reference translation. The baseline system did not capture this and translated, explicitly, both connectives: *since* – *da* and *then* – *dann*. There are two problems with such a translation: *since* is translated to *da* which in DE clearly signals a CAUSAL discourse relation and is wrong in this context. Even if *since* would have been translated to the TEMPORAL *seit*, the following *dann* would not sound fluent. The correct solution therefore is to omit, or implicitate the connective *then* and to produce one connective only, i.e. *seitdem*, as the `swd-4` learned correctly via sparse features and the TEMPORAL label for *since*.

Example 2 : For example sentence 2. we rated equal readability for the `swd-3` and baseline translations, as they both produced the acceptable *so weit* as a DE paraphrase for the EN connective *as far as*. In the human reference translation, *as far as* has safely been implicitated in DE via different word order (the verb *verwendet* (EN *use*) can replace the EN *went to use*), i.e. there is no lack of meaning or readability to observe. That *as far as* could have been implicitated in this way was neither learned by the baseline nor the `swd-3` model, that would have had a feature for *as far as*. We checked in the feature weights and indeed found a negative probability value for *as far as* (-0.0042), which indicates that this connective is rarely implicitated.

Example 3 : In example sentence 3. the EN connectives are *whereas* and *previously*. This was one of the cases where the `swd-4` model degraded the readability of the translation, by deleting the wrong connective of the two: the model omits *previously* and translates *whereas* with the correct DE connective *während*. However, by omitting *previously*, the temporal information that the contract only held for one year up to now is lost and not recoverable. The baseline SMT system here scores better for readability, as it translates both connectives and preserves the contrastive (*whereas*) and temporal relation (*previously*). In the humane reference translation, remarkably, it is the CONTRAST discourse relation that is omitted (no translation for *whereas*). The TEMPORAL relation is translated in DE with the paraphrase bis jetzt). Together with word order and the reinforcing expressions of *immer nur* (EN: *always only*), the CONTRAST relation can here implicitly be inferred by a human reader.

When overall counting for 50 `swd-3` and 50 `swd-4` translations, we observe the following percentages of better, equal and worse translations in Table 8.8. We additionally scored for BLEU, in each of the 50 evaluated translations.

These experiments confirm that considering a large number of connectives for translation (as in the `swd-3` model) may not be useful, as most of them are actually *unproblematic* for a baseline system, due to their frequency and unambiguity, at least for the EN/FR/DE/ translation directions considered in this thesis. When it comes to highly ambiguous connectives however, again, labeling them prior to translation with the discourse relation they signal helps finding not only more correct explicit translations, but can also help finding cases where it is more 'natural' or fluent for a target language to omit or paraphrase the connectives in the

Model	$\Delta Read.$ (%)			BLEU scores	
	+	=	−	Model	Baseline
swd-3	44	8	48	10.08	8.16
swd-4	60	6	34	10.76	8.30

Table 8.8: Manual evaluation scores on 50 translations for readability ($\Delta Read.$) for the SMT models swd-3 and swd-4 that made use of sparse lexical features to predict how likely it is to delete a connective. Both models increase the BLEU score compared to the baseline, due to implicitation that makes the model translations more similar to the human reference. The manual readability counts show that overall model swd-4 with labeled connectives helped to improve about 26% of the translations, while model swd-3 decreased translation quality by 4%.

translation process. Moreover, the labeling with classifiers and the subsequent translation modeling with sparse lexical features provides a completely automated setting with which the human level of implicitation for connectives can be learned for SMT.

9 Conclusions and perspectives

In this thesis, we have addressed a number of translation problems regarding linguistic phenomena that are established at the discourse-level of texts, beyond single sentences. As current SMT algorithms cannot handle these due to their constraint to translate on a sentence-by-sentence basis, text-level phenomena such as discourse connectives and verb tenses are often wrongly captured by baseline systems. The methods discussed in this thesis have lead to fully operational and automated SMT system pipelines, in which classifiers trained through machine learning are used to label the needed discourse relations and verb tenses onto the lexical material of the source language, prior to SMT.

9.1 Conclusions

As semantics and discourse only very recently have been addressed for SMT, we first started by cross-linguistic, contrastive corpus analyses of discourse connectives and verb tenses, revealing their considerable ambiguity and divergence between languages. With both types of inter-sentential relations, translation errors can go as far as misleading readers regarding the argumentative structure a text conveys, or the ordering of events a text describes.

The set of available connectives in a source language does not map one-to-one to the set of the target language. In addition, the ambiguity of a source language connective might or might not be preserved in the target language, where either more, fewer, or even no connectives can be available to express the corresponding discourse relation. Similarly, the verb tense system of two languages can differ considerably, i.e. the available source language tenses do usually not map to target tenses in a one-to-one fashion, but rather, depending on the context and the current discourse, various target forms may be appropriate in translation.

The contrastive corpus analyses helped us to find the discourse connectives and verb tenses that are most problematic in translation. From previous work, we knew, for example, that most of the 100 English PDTB connectives are actually not ambiguous and are therefore rather straightforward to translate. A small subset however, can be highly ambiguous, comprising

connectives such as *although, however, meanwhile, since, though, while* or *yet*, which can signal up to seven discourse relations, as observed in the datasets we examined. Having reimplemented a series of state-of-the-art syntactic features and extending them by newly found features such as dependency tags, WordNet similarity scores and antonyms, baseline translations, polarity values and neighboring discourse relations, we obtained specific classifiers for each connective, reaching F1 scores in ranges of 0.7 to 0.9. In manual and automatic annotation we considered only those senses (signaled by these connectives) that are at the granularity level that is necessary to find their correct target language equivalents.

The merits of the newly proposed features were confirmed by directly using the classifiers to annotate connective occurrences in large portions or the entire Europarl corpus for the EN/FR, EN/DE, EN/IT language pairs and the UN corpus for EN/AR, as training data for SMT systems. By considering various methods to make use of the classifiers' labels for machine translation, such as connective-label concatenation, system combination and factored translation models, we were able to gain improvements in BLEU scores of about 0.2-0.4, while the translation of connectives was improved by up to 10%, as found through manual evaluation or by using a new metric named ACT that considers connectives and possible target language equivalents.

With a new SMT tuning method and sparse lexical features, we were able to further show that the labels of discourse relations can help implicitating target language connectives in the correct places and more accurately than a system that was tuned in similar manner, but had only unlabeled connectives as features.

To demonstrate the generality of our approach, we implemented two classifiers for a second discourse-level phenomenon: verbal tenses. We focused first on the EN Simple Past, which has two usages (indicating either events or state of affairs in the past) that have to be translated into at least three different tenses in FR (*Passé Composé, Imparfait, Passé Simple*). The first classifier we designed relies on specific features for temporality that have been extracted: besides constituent features and the context of the verb phrase, temporal connectives are especially helpful to point to the correct ordering of events. The classifier is able to detect whether the context of each EN Simple Past tense verb is a narrative one or not with an F1 score of 0.72. By applying this classifier prior to a factored translation model, the translation of the EN Simple Past tense was improved in 9% of all cases, as shown by the manual evaluation of tense, lexical verb choice and verb phrase correctness on 200 test instances.

We extended this idea, given the fact that there are many more verb tense mismatches for the EN/FR language pair than the one involving Simple Past, especially between the available past tenses of the two languages, but also when there is no EN tense equivalent, as with the FR *Subjonctif*, for example, or the FR *Présent* that can be a valid translation of all EN tenses considered here. We have exploited a large, high-precision resource in which all EN verbs are labeled with the FR tense they were translated to in the reference to train a tense-aware, factored SMT system that improves verb tense translation by up to 25% (while also maintaining correctness of lexical choice and person/number agreement) and by 0.5 BLEU points overall

quality using the oracle labels. Then, we built a verb tense predictor (for EN to FR translation) that uses a richer feature set than the narrativity classifier: dependency tags and semantic role labeling, in addition to TimeML, syntactic and discourse connective features.

In its best configuration, the classifier reaches a performance of 0.83 F1 score overall, although it is biased toward frequent present tense usage. We therefore added two features for the FR tenses of *Imparfait* and *Subjonctif*, both difficult to predict and translate from EN. Classification for these two tenses is only slightly improved by 0.01 to 0.04 F1 score, but the effect on tense translation is however noticeable: from 23.61 to 24.30 BLEU points when considering sentences with labeled tenses only. The effect is also confirmed by manual evaluation.

The machine translation experiments described in this thesis confirm the validity of the initial hypothesis: by inserting linguistic information at the discourse level of source texts, their automatic translation can be improved. The translation models can be trained on texts with either manually annotated or automatically classified discourse connectives and verb tenses. Both types of discourse information have a considerable effect on the readability and the perception of a translated text by human readers in terms of correct argumentative text structures and the correct temporal ordering of the events described.

The thesis is a timely contribution to a field that just started to consider the importance of discourse-level features for statistical MT. The focus of other work has however been on lexical consistency and content words, whereas our work shed light on the importance of the correct translation of more functional categories. We however not only aimed at translation quality improvements but also provided features and methods in order to more correctly classify these discourse phenomena automatically, which are NLP tasks of their own and have a potential influence on other applications as well, as discussed in the perspectives below. The present thesis also confirms what has been found in previous research work: building specific classifiers, focusing on single connective types, leads to better performance and robustness than trying to jointly classify for several discourse relations.

The main approach taken, i.e. the coupling of manual or automatic annotations with factored translation models, has the advantage that the influence of correct labeling has a direct and measurable effect on translation quality: translations based on manually labeled discourse features score the highest, and for automatic classifiers, the higher the performance the better the resulting translations. The quality of the translations is, as we have shown as well, heavily dependent on the data used to train, tune and test an SMT system: the higher the proportion of "easy to classify" connectives in the corresponding dataset, the higher the resulting performance and, as a consequence, the higher the resulting translation quality when combining the systems. Moreover, factored translation models are influenced by parameter choice and the amount of data used: the larger the training corpus, the smaller the effect of using linguistic labels when comparing to a standard baseline phrase-based translation model.

9.2 Perspectives

The combination of systems proposed in this thesis can appear to be quite a complex series of processes: potentially costly manual annotation, followed by classifier training based on potentially noisy and computationally demanding feature extraction, followed by SMT training and tuning with high computational costs, resulting in sometimes modest improvements in translation quality.

As a very first step in future work, therefore, one might want to conceptually re-think the 'dual' system architecture of classification followed by translation. Recent progress in SMT decoding allows for giving more weight to the previous translations and features up to the document-level, thanks to the Docent decoder Hardmeier et al. [2012]. This would likely be beneficial to the translation of connectives and verb tenses, but also of referring expressions such as noun phrases and pronouns, as the corresponding feature functions in such translation models would help to increase the weight of phrase-pairs that better fit the context of previously translated units. Moreover, this approach would have the advantage to simplify processing, as no classifiers with time-consuming feature extraction and no disambiguation labels would be needed, and the scoring of appropriate discourse unit translations would directly take place in the translation model.

We made a preliminary test to extend Docent by implementing a new feature function for a document-level feature concerning discourse connectives. In order to move away from the sentence level and judgments whether certain discourse connectives should be translated or not, a first experiment is to count, during decoding, the source language connectives that do not have, in their target phrase pair, an explicit translation as stored in the connective dictionary described in Chapter 8. If we count all those entries in the phrase table that appear during decoding we can: (a) give a score to the whole document (for the amount of implicitly translated connectives), and (b) decrease the score of a phrase pair that actually consists of an explicit target language connective translation. By doing this, we achieve a similar BLEU score over the nt2010+2012 WMT test data (used several times in this thesis, see Chapters 7 and 8), when compared to a baseline Docent system, not considering the connective feature. When examining the translation, we see that the feature function behaves as expected and does not generate a connective in the translations where it can be implicitated. Further experiments into this direction would likely advance discourse unit translation.

With the advancements of current SMT paradigms, the hierarchical, syntactical tree-to-string, string-to-tree and tree-to-tree models using grammatical rule implementations or syntactic parsing could be extended with discourse parsing, i.e. trees over entire paragraphs rather than sentences only, as it has been already proposed by Marcu et al. [2000]. This however requires considerable improvements in the performance of discourse parsers (current accuracy levels are at 40-70%), given that syntactical SMT does most often not reach the quality level of phrase-based MT even though syntactic parsers have a higher performance (80-90% accuracy) than discourse ones.

We have additionally shown that new SMT decoding and tuning methods, such as sparse lexical features, can lead to more fluent and natural sounding target text. Such models could be investigated further in future work, as we only have tried here a few configurations and only for cases where EN connectives should be omitted when translating to DE. These features are likely to help the opposite case as well: when a connective should be inserted (explicitated) in the target text. This is certainly more difficult to achieve than deleting a word that is realized on the surface, because it requires generating and inserting a lexical word form without any source surface form as a basis.

More generally, the divergence or mismatches of source and target languages could, in future work, be addressed from a broader perspective than discourse markers. Recent research in monolingual corpora and in MT have brought up the idea of studying other aspects of textual meaning and/or coherence through the notion of *paraphrasing*. Most often considered when having to translate out-of-vocabulary noun phrases, paraphrasing source text prior to translation to augment the training data could, as we have pointed to in Chapter 8, be utilized to translate discourse units as well.

Paraphrasing in SMT is usually considered in the cases when, for a given source phrase, no target phrase can be found in the phrase table. One way to address this, is by finding paraphrases by pivoting through phrases in another language. The target language translations of an EN phrase are identified, all occurrences of those target phrases are found, and all EN phrases that they translate back to are treated as potential paraphrases of the original EN phrase (Callison-Burch et al. [2006]). Research into paraphrasing and entailment methods are worthwhile, especially for the problem of translationese and large source/target divergencies and mismatches. Re-formulating the source prior to translation or finding alternative target phrases is likely to lead to translation candidates that are more fluent and readable. As we have shown in this thesis, discourse elements are especially prone to be affected by translationese and paraphrasing.

We believe that these approaches represent timely and necessary investigations in SMT in order to make progress toward fully automatic, high-quality MT. Discourse modeling will have to be considered in future work, although improvements for syntactical and semantic models are still needed. If translations fail because of wrong word/constituent order or because of mismatching semantic concepts between source and target language, the readability of the text can be as negatively affected as it would be through wrong argumentative structuring of sentences and paragraphs at the discourse level. This thesis has provided early research and reproducible methods, which should be helpful for future work that tries to advance SMT toward coherent and well-structured, human-readable target text.

A Appendix

List of temporal markers used as features for verb tense disambiguation (A.1, see also Chapter 6) and tables of English (A.2), French (A.3) and German (A.4) connectives considered for implicitation and explicitation when translating from English (Chapter 8).

Table A.1: Manually compiled list of temporal markers used for verb tense disambiguation: 's' denotes synchrony and 'a' asynchrony.

Temporal marker	**Sense**
after	a
months after	a
month after	a
years after	a
year after	a
weeks after	a
week after	a
days after	a
day after	a
hours after	a
hour after	a
minutes after	a
minute after	a
immediately after	a
even after	a
only after	a
shortly after	a
soon after	a
afterwards	a

afterward	a
and	a/s
as	a/s
as long as	s
as soon as	a/s
before	a
months before	a
month before	a
years before	a
year before	a
weeks before	a
week before	a
days before	a
day before	a
hours before	a
hour before	a
minutes before	a
minute before	a
before and after	a
but	a
by then	a/s
earlier	a
finally	a
if	s
if and when	s
in the end	a
in turn	a
later	a
meantime	s
meanwhile	s
much as	s
next	a
now that	a/s
once	a
previously	a
separately	s
simultaneously	s
since	a
still	a
then	a/s

thereafter	a
till	a
until	a
when	a/s
when and if	a
while	s
yet	a

Table A.2: English connectives with a frequency above 20 in the PDTB. Also listed are the level-2 majority relations with the number of tokens out of the total tokens of the connective in the PDTB (counts including the majority relation being part of a composite sense tag). *For some connectives there is no level-2 majority because some instances have only been annotated with level-1 senses. We did not consider the connectives *and* and *or* (too many non-connective occurrences for automatic detection).

EN connective	**Majority relation**	**Tokens**
after	ASYNCHRONY	575/577
also	CONJUNCTION	1735/1746
although	CONTRAST	*157/328
as	SYNCHRONY	543/743
as a result	CAUSE	78/78
as if	CONCESSION	*4/16
as long as	CONDITION	20/24
as soon as	ASYNCHRONY	11/20
because	CAUSE	854/858
before	ASYNCHRONY	326/326
but	CONTRAST	2427/3308
by contrast	CONTRAST	27/27
even if	CONCESSION	*41/83
even though	CONCESSION	72/95
finally	ASYNCHRONY	*14/32
for example	INSTANTIATION	194/196
for instance	INSTANTIATION	98/98
however	CONTRAST	355/485
if	CONDITION	1127/1223
in addition	CONJUNCTION	165/165

indeed	CONJUNCTION	54/104
in fact	RESTATEMENT	*39/82
instead	ALTERNATIVE	109/112
in turn	ASYNCHRONY	20/30
just as	SYNCHRONY	13/14
later	ASYNCHRONY	90/91
meanwhile	SYNCHRONY	148/193
moreover	CONJUNCTION	100/101
nevertheless	CONCESSION	*19/44
nonetheless	CONCESSION	17/27
now that	CAUSE	20/22
once	ASYNCHRONY	78/84
on the other hand	CONTRAST	35/37
otherwise	ALTERNATIVE	22/24
previously	ASYNCHRONY	49/49
separately	CONJUNCTION	73/74
since	CAUSE	104/184
so that	CAUSE	31/31
still	CONCESSION	83/190
then	ASYNCHRONY	312/340
therefore	CAUSE	26/26
though	CONCESSION	*156/320
thus	CAUSE	112/112
unless	ALTERNATIVE	94/95
until	ASYNCHRONY	140/162
when	SYNCHRONY	594/989
while	CONTRAST	455/781
yet	CONTRAST	53/101

Table A.3: French connectives taken from LexConn Roze et al. [2010] with RST-like discourse relation labels as used in the original work.

FR connective	**Majority relation**
dans ce cas	CONSEQUENCE
tout à coup	NARRATION
surtout que	EXPLANATION
d'un autre côté	CONTRAST

tandis que	CONTRAST
par contre	CONTRAST
de toute façon	DETACHMENT
à condition de	CONDITION
mais	CONTRAST
à fin que	GOAL
malgré le fait que	CONCESSION
puisque	EXPLANATION
avant que	NARRATION
ou bien	ALTERNATION
au moment de	BACKGROUND
après que	FLASHBACK
quant à	BACKGROUND
à condition que	CONDITION
auparavant	FLASHBACK
malgré tout	VIOLATION
par comparaison	CONTRAST
surtout	CONTINUATION
tout d'abord	ELABORATION
en revanche	CONTRAST
de fait	EXPLANATION
plutôt que	BACKGROUND
bien que	CONCESSION
tant que	CONDITION
de même que	PARALLEL
de sorte que	RESULT
encore	VIOLATION
avant de	NARRATION
pendant que	CONTRAST
quoique	CONCESSION
en ce cas	CONSEQUENCE
en comparaison	CONTRAST
même si	CONCESSION
en fait	BACKGROUND
de la même façon	PARALLEL
néanmoins	VIOLATION
dès lors	RESULT
par conséquent	RESULT
si	CONDITION
alors que	CONTRAST

par ailleurs	CONTINUATION
au cas ou	CONDITION
en gros	SUMMARY
depuis	NARRATION RESULT
encore que	CONCESSION
au contraire	CONTINUATION
effectivement	EVIDENCE
à ce moment là	CONSEQUENCE
simultanément	PARALLEL
autrement	ALTERNATION
ensuite	NARRATION
cependant	VIOLATION
en même temps	VIOLATION
car	EXPLANATION
sans que	BACKGROUND
à partir du moment ou	CONDITION
en outre	CONTINUATION
et puis	CONTINUATION
après tout	EXPLANATION
au lieu	VIOLATION
à part ça	VIOLATION
tout en	CONCESSION
depuis que	FLASHBACK EXPLANATION
à moins de	ALTERNATION
d'ailleurs	EVIDENCE
quand	BACKGROUND INVERSE
en tout cas	DETACHMENT
de même	PARALLEL
finalement	NARRATION
étant donné que	EXPLANATION
ainsi que	CONTINUATION
ainsi	RESULT
jusqu'à	RESULT
lorsque	BACKGROUND INVERSE
avant	FLASHBACK
quoi qu'il en soit	DETACHMENT
enfin	CONTINUATION
après	NARRATION
avant même de	BACKGROUND
en plus	CONTINUATION

donc	RESULT
pourtant	VIOLATION
dès que	FLASHBACK EXPLANATION
en effet	EXPLANATION
plus que	BACKGROUND
tout de même	VIOLATION
alors même que	CONCESSION
en attendant	DETACHMENT
jusqu'à ce que	GOAL RESULT
plus tard	NARRATION
malgré que	CONCESSION
par exemple	ELABORATION
vu que	EXPLANATION
sachant que	EXPLANATION
outre que	CONTINUATION
quand même	VIOLATION
parce que	EXPLANATION
en conséquence	RESULT
pour que	GOAL
sauf que	VIOLATION
or	CONTINUATION
en même temps que	CONCESSION
afin de	GOAL
comme	EXPLANATION
dès lors que	CONDITION
du coup	RESULT
une fois que	FLASHBACK
de manière que	GOAL
d'autre part	CONTINUATION
plutôt	CONTINUATION
puis	NARRATION
de toute manière	DETACHMENT
cela dit	VIOLATION
d'un coup	NARRATION
dans la mesure où	EXPLANATION
également	PARALLEL
dans ce cas là	CONSEQUENCE
de la même manière	PARALLEL
sinon	ALTERNATION
au moment où	BACKGROUND

a moins que	ALTERNATION
dans le cas où	CONDITION
avant même que	BACKGROUND
sans doute	VIOLATION
avant tout	BACKGROUND
autant que	EXPLANATION
c'est pourquoi	RESULT

Table A.4: German connectives taken from DimLex (Stede and Umbach [1998]) with RST-like discourse relation labels as used in the original resource.

German connective	**Majority relation**
soweit	NOTYET
allerdings	CONCESSION
da	JOINT
dann	SEQUENCE
hiernach	SEQUENCE
dadurch	ELABORATION
weder	NOTYET
anstelle dessen	ASYMMETRIC CONTRAST
dagegen	CONTRAST
ansonsten	CONTRAST
bevor	SEQUENCE
also	CAUSE
sodann	SEQUENCE
vorausgesetzt, dass	CONDITION
obgleich	CONCESSION
obendrein	ELABORATION
weiterhin	ELABORATION
ausserdem	CIRCUMSTANCE
infolgedessen	CAUSE
sonst	NOTYET
inzwischen	SEQUENCE
folglich	CAUSE
andererseits	CONTRAST
anstatt	CONTRAST
weshalb	CAUSE

ebenso wenig	ELABORATION
aufgrund	CAUSE
dennoch	CONCESSION
nebenher	JOINT
hierdurch	CAUSE
dabei	ELABORATION
desgleichen	ELABORATION
zuvor	SEQUENCE
ausser, dass	CONCESSION
hingegen	CONTRAST
womit	NOTYET
allein	CONTRAST
darauf	SEQUENCE
sodass	CAUSE
sofern	CONDITION
nur, dass	CONCESSION
soeben	NOTYET
deshalb	CAUSE
dessen ungeachtet	CONCESSION
zugleich	JOINT
obwohl	CONCESSION
später	SEQUENCE
beispielsweise	ELABORATION
zusätzlich	ELABORATION
wegen	CAUSE
ohnehin	ELABORATION
vorher	NOTYET
nichtsdestoweniger	CONCESSION
trotzdem	CONCESSION
ferner	ELABORATION
hierauf	SEQUENCE
statt	NOTYET
nämlich	ELABORATION
ob	NOTYET
wonach	SEQUENCE
während	JOINT
weswegen	CAUSE
darüber hinaus	ELABORATION
zumal	CAUSE
nachdem	PRECONDITION

so oder so	ELABORATION
trotz	CONCESSION
wie	SEQUENCE
deswegen	CAUSE
wenn	NOTYET
des weiteren	ELABORATION
beziehungsweise	ELABORATION
ergo	CAUSE
inwieweit	ELABORATION
unterdessen	JOINT
zunächst	SEQUENCE
ohnedies	ELABORATION
falls	CONDITION
doch	ELABORATION
infolge	CAUSE
woraufhin	SEQUENCE
ausser wenn	CONCESSION
zum Beispiel	ELABORATION
warum	ELABORATION
daher	CAUSE
demgegenüber	CONTRAST
ebenfalls	ELABORATION
wogegen	CONTRAST
mithin	CAUSE
seit	NOTYET
überdies	ELABORATION
denn	CAUSE
obzwar	CONCESSION
entgegen	CONTRAST
inwiefern	ELABORATION
zudem	ELABORATION
seither	NOTYET
zwar	CONCESSION
ausser	CONCESSION
so	ELABORATION
aus diesem Grund	CAUSE
daraufhin	SEQUENCE
allenfalls	ELABORATION
darum	CAUSE
wenngleich	CONCESSION

als	CIRCUMSTANCE
somit	CAUSE
weil	CAUSE
indem	MEANS
insofern	MEANS
währenddessen	NOTYET
gleichwohl	CONCESSION
andernfalls	CONTRAST
dafür	CONTRAST
danach	SEQUENCE
worauf	SEQUENCE
sobald	SEQUENCE
seitdem	NOTYET
obschon	CONCESSION
aber	CONTRAST
sondern	CONTRAST
vielmehr	ASYMMETRIC CONTRAST
wenn auch	CONCESSION
auch wenn	CONCESSION

Bibliography

Bas Aarts. *Oxford Modern English Grammar*. Oxford University Press, 2011.

Amal Alsaif. *Human and Automatic Annotation of Discourse Relations for Arabic*. PhD thesis, University of Leeds, 2012.

C.F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Lingusitics (ACL)*, pages 86–90, Montreal, Canada, 1998.

Kathryn Baker, Bonnie Dorr, Michael Bloodgood, Chris Callison-Burch, Nathaniel Filardo, Christine Piatko, Lori Levin, and Scott Miller. Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, 38(2):411–438, 2012.

Satanjeev Banerjee and Ted Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 117–171. Springer, Berlin/Heidelberg, Germany, 2002.

Marco Baroni and Silvia Bernardini. A New Approach to the Study of Translationese: Machine-Learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3), 2005.

Bergljot Behrens and Cathrine Fabricius-Hansen. Translation Equivalents as Empirical Data for Semantic/Pragmatic Theory. In K. Jaszczolt and J. Turner, editors, *Meaning through Language Contrast*, pages 463–477. Benjamins, Amsterdam, Netherlands, 2003.

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, 2007.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Rober L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 17–24, New York, NY, 2006.

Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22:249–254, 1996.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. The RST Discourse Treebank. In *The Linguistic Data Consortium*, Philadelphia, PA, 2002.

Marine Carpuat. One Translation per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW)*, pages 19–27, Singapore, 2009.

Marine Carpuat and Michel Simard. The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pages 442–449, Montreal, Canada, 2012.

Marine Carpuat and Dekai Wu. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 61–72, Prague, Czech Republic, 2007.

Bruno Cartoni and Thomas Meyer. Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR, 2011.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Using the Europarl Corpus for Cross-linguistic Research. *Belgian Journal of Linguistics*, 27:23–42, 2013a. doi: http://dx.doi.org/10.1075/bjl.27.02car.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation-spotting Technique. *Dialogue & Discourse*, 4(2):65–86, 2013b.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 33–40, Prague, Czech Republic, 2007.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Pi-Chuan Chang, Dan Jurafsky, and Christopher D. Manning. Disambiguating 'DE' for Chinese-English Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece, 2009.

Eugene Charniak and Mark Johnson. Coarse-to-fine n-best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 173–180, Ann Arbor, MI, 2005.

Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 427–436, Montreal, Canada, 2012.

David Chiang. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, MI, 2005.

David Chiang. Hope and Fear for Discriminative Training of Statistical Translation Models. *Journal of Machine Learning Research*, 13(1):1159–1187, 2012.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. Learning Morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies)*, Portland, OR, 2011.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research (JMLR)*, 11:2461–2505, 2011.

Francis Corblin and Henriëtte de Swart. *Handbook of French Semantics*. CSLI Publications, Stanford, CA, 2004.

Laurence Danlos and Charlotte Roze. Traduction (automatique) des connecteurs de discours. In *Actes de la 18è Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, 2011.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB : French Discourse Tree Bank. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 471–478, Grenoble, France, 2012.

Pascal Denis and Benoît Sagot. Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-art POS Tagging with less Human Effort. In *Proceedings of PACLIC 2009 (23rd Pacific Asia Conference on Language, Information and Computation)*, pages 110–119, Hong Kong, China, 2009.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, UK, 2011.

Bonnie J. Dorr. A Parameterized Approach to Integrating Aspect with Lexical Semantics for Machine Translation. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 257–264, Newark, DE, 1992.

Bonnie J. Dorr and Terry Gaasterland. Selecting Tense, Aspect, and Connecting Words in Language Generation. In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1299–1307, Montreal, Canada, 1995.

David duVerle and Helmut Prendinger. A Novel Discourse Parser Based on Support Vector Machine Classification. In *Proceedings of ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP)*, pages 665–673, Singapore, 2009.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A Decoder, Alignment, and Learning framework for finite-state and context-free translation models. In *Proceedings of the 48th Conference of the Association for Computational Linguistics (ACL), System Demonstrations*, pages 7–12, Uppsala, Sweden, 2010.

Helge Dyvik. A Translational Basis for Semantics. In Stig Johansson and Signe Okselfjell, editors, *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, pages 51–86. Rodopi, Amsterdam, Netherlands, 1998.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of ACL 2012 (50th Annual Meeting of the Association for Computational Linguistics*, pages 115–119, Jeju, Republic of Korea, 2012.

Robert Elwell and Jason Baldridge. Discourse Connective Argument Identification with Connective Specific Rankers. In *Proceedings of ICSC 2008 (2nd IEEE International Conference on Semantic Computing)*, pages 198–205, Santa Clara, CA, 2008.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, Brisbane, Australia, 2008.

Anita Gojun and Alexander Fraser. Determining the Placement of German Verbs in English-to-German SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 726–735, Avignon, France, 2012.

Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. N-Gram-Based Tense Models for Statistical Machine Translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 276–285, Jeju Island, Korea, 2012.

Cristina Grisot and Bruno Cartoni. Une description bilingue des temps verbaux: étude contrastive en corpus. *Nouveaux cahiers de linguistique française*, 30:101–117, 2012.

Cristina Grisot and Thomas Meyer. Cross-linguistic Annotation of Narrativity for English/French Verb Tense Disambiguation. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

Liane Guillou. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of EACL 2012 Student Research Workshop (13th Conference of the European Chapter of the ACL*, pages 1–10, Avignon, France, 2012.

Nizar Habash and Owen Rambow. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 573–580, Ann Arbor, Michigan, 2005.

Barry Haddow. Acquiring a Disambiguation Model For Discourse Connectives. Master's thesis, School of Informatics, University of Edinburgh, Scotland, UK, 2005.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank 2.0. Technical report, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2011. URL https://ufal.mff.cuni.cz/pdt2.0/.

Najeh Hajlaoui and Andrei Popescu-Belis. Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Samos, Greece, 2013.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, 2009.

M. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

Sandra Halverson. Connectives as a Translation Problem. In H. et al. (Eds.) Kittel, editor, *Encyclopedia of Translation Studies*, pages 562–572. Walter de Gruyter, Berlin/New York, 2004.

Christian Hardmeier. Discourse in Statistical Machine Translation. *DISCOURS*, 11:1–29, 2013. URL http://discours.revues.org/8726.

Christian Hardmeier and Marcello Federico. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2010.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju, Korea, 2012.

James Henderson, Paola Merlo, Gabriele Musillo, and Ivan Titov. A Latent Variable Model of Synchronous Parsing for Syntactic and Semantic Dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL)*, pages 178–182, Manchester, UK, 2008.

Hugo Hernault, Danushka Bollegala, and Ishizuka Mitsuru. Semi-supervised Discourse Relation Classification with Structural Learning. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 340–352, Tokyo, Japan, 2011.

Stéphane Huet, Julien Bourdaillet, and Philippe Langlais. Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, 2009.

Iustina Ilisei, Diana Inkpen, Gloria Pastor Corpas, and Ruslan Mitkov. Identifcation of Translationese: A Machine Learning Approach. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg, Germany, 2010.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. FEMTI: Creating and Using a Framework for MT Evaluation. *Proceedings of MT Summit IX, New Orleans, LA*, pages 224–231, 2003.

K. S. Kipper. *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. Phd thesis, University of Pennsylvania, 2005.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, Cambridge UK, 2010.

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, pages 868–876, Prague, Czech Republic, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-based Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic, 2007.

Sudheer Kolachina, Rashmi Prasad, Dipti Misra Sharma, and Aravind Joshi. Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Instanbul, Turkey, 2012.

Moshe Koppel and Noam Ordan. Translationese and its Dialects. In *Proceedings of ACL-HLT 2011 (49th Annual Meeting of the ACL: Human Language Technologies*, pages 1318–1326, Portland, OR, 2011.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *The Journal of Machine Learning Research*, 8:693–723, 2001.

Sara Laviosa-Braithwaite. *The English Comparable Corpus (ECC): A Resource and Methodology for the Empirical Study of Translation.* Phd thesis, University of Manchester, UK, 1996.

Ronan Le Nagard and Philipp Koehn. Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, Uppsala, Sweden, 2010.

Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. Generating Discourse Structures for Written Text. In *Proceedings of Coling*, pages 329–335, Geneva, Switzerland, 2004.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving Statistical Machine Translation by Adapting Translation Models to Translationese. *Computational Linguistics*, 39(4): 999–1023, 2013. doi: 10.1017/S1351324912000307.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 343–351, Singapore, 2009.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20(2):151–184, 2014. doi: 10.1017/S1351324912000307.

Diane J. Litman. Classifying Cue Phrases in Text and Speech Using Machine Learning. In *Proceedings of the Annual Meeting of the American Association for Artificial Intelligence*, pages 806–813, Seattle, WA, 1994.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

H. Lüngen, H. Lobin, M. Bärenfänger, M. Hilbert, and C. Puskas. Text parsing of a Complex Genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB*, Bansko, Bulgaria, 2006.

Jianjun Ma, Degen Huang, Haixia Liu, and Wenfeng Sheng. POS Tagging of English Particles for Machine Translation. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 57–63, Xiamen, China, 2011.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988.

Christopher Manning and Dan Klein. Optimization, MaxEnt Models, and Conditional Estimation without Magic. In *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan, 2003.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization.* A Bradford Book. The MIT press, Cambridge, MA, London, UK, 2000.

Daniel Marcu, Lynn Carlson, and Maki Watanbe. The Automatic Translation of Discourse Structures. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17, Philadelphia, PA, 2000.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Robert Martin. *Temps et aspect: essai sur l'emploi des temps narratifs en moyen français.* Klincksieck, Paris, France, 1971.

Thomas Meyer. Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation. In *Proceedings of ACL-HLT 2011 (49th Annual Meeting of the ACL: Human Language Technologies), Student Session*, pages 46–51, Portland, OR, 2011.

Thomas Meyer and Lucie Poláková. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 43–50, Sofia, Bulgaria, 2013.

Thomas Meyer and Andrei Popescu-Belis. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting*

Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra), pages 129–138, Avignon, FR, 2012.

Thomas Meyer and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 19–26, Sofia, Bulgaria, 2013.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, pages 194–203, Portland, OR, 2011.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA, 2012.

Thomas Meyer, Cristina Grisot, and Andrei Popescu-Belis. Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 33–42, Sofia, Bulgaria, 2013.

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Transactions of Audio, Speech, and Language Processing*, submitted, 2014.

Rada Mihalcea and Andras Csomai. SenseLearner: Word Sense Disambiguationfor All Words in Unrestricted Text. In *Proceedings of the ACL 2005 Interactive Poster and Demonstration Sessions*, pages 53–56, Ann Arbor, MI, 2005.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38 (11):39–41, 1995.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, Barcelona, Spain, 2005.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind K. Joshi. Sense Annotation in the Penn Discourse Treebank. *Lecture Notes in Computer Science*, 4919:275–286, 2008.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. Source-language Entailment Modeling for Translating Unknown Terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 791–799, Suntec, Singapore, 2009.

Ruslan Mitkov. Introduction: Special Issue on Anaphora Resolution in Machine Translation and Mulitlingual NLP. *Machine Translation*, 14:159–161, 1999.

Rebecca Nesson, Stuart M. Shieber, and Alexander Rush. Induction of Probabilistic Synchronous Tree-insertion Grammars for Machine Translation. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, MA, 2006.

S. Nirenburg, H.L. Somers, and Y. Wilks. *Readings in Machine Translation*. A Bradford book. MIT Press, 2003.

Joakim Nivre. An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160, Tokyo, Japan, 2003.

Dick Noël. Translations as Evidence for Semantics: An Illustration. *Linguistics*, 41:757–785, 2003.

Michal Novak, Anna Nedoluzhko, and Ždeněk Žabokrtský. Translation of 'It' in a Deep Syntax Framework. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 51–59, Sofia, Bulgaria, 2013.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, 2003.

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Sylwia Ozdowska. Donnees bilingues pour la TAS francais-anglais: impact de la langue source et direction de traduction originales sur la qualite de la traduction. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France, 2009.

M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105, 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, 2002.

Gary Patterson and Andrew Kehler. Predicting the Presence of Discourse Connectives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 914–923, Melbourne, Australia, 2013.

Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, Laure Vieu, and Antoine Widlöcher. ANNODIS: une approche outillée de l'annotation de structures

discursives. In *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Paris, France, 2009.

Emily Pitler and Ani Nenkova. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP), Short Papers*, pages 13–16, Singapore, 2009.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. Easily Identifiable Discourse Relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Companion Volume: Posters*, pages 87–90, Manchester, UK, 2008.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP)*, pages 683–691, Singapore, 2009.

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan, 2013.

Andrei Popescu-Belis and Sandrine Zufferey. Contrasting the Automatic Identification of Two Discourse Markers in Multiparty Dialogues. In *Proceedings of the 8th SIGdial Meeting on Discourse and Dialog*, pages 10–17, Antwerp, Belgium, 2007.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, The PDTB Research Group, http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf, December 17, 2007.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Morocco, 2008.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1023–1031, Beijing, China, 2010.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The Biomedical Discourse Relation Bank. *BMC Bioinformatics*, 12(188):1–18, 2011.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. The TimeBank Corpus. In *Proceedings of Corpus Linguistics Conference*, pages 647–656, Lancaster, UK, 2003.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Pearson Longman, Harlow, UK, 1986.

Alexandre Rafalovitch and Robert Dale. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *In Proceedings of MT Summit XII*, pages 292–299, Ontario, Canad, 2009.

Charlotte Roze, Laurence Danlos, and Phillippe Muller. LEXCONN: a French Lexicon of Discourse Connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD)*, pages 114–125, Moissac, France, 2010.

Tanja Samardzic, Lonneke van der Plas, Goljihan Kashaeva, and Paola Merlo. Variation in Verbal Predicates in English and French. *Proceedings of GG@G: Generative Grammar in Geneva*, 6:109–135, 2010.

Lucia Silva. Fine-Tuning in Brazilian Portuguese-English Statistical Transfer Machine Translation: Verbal Tenses. In *Proceedings of the NAACL HLT Student Research Workshop*, pages 58–63, Los Angeles, CA, 2010.

Michel Simard. Translation Spotting for Translation Memories. In *Proceedings of the HLT-NAACL workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond*, pages 65–72, Edmonton, Canada, 2003.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, MA, 2006.

Radu Soricut and Daniel Marcu. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156, Edmonton, CA, 2003.

Wilbert Spooren and Ted Sanders. The Acquisition Order of Coherence Relations: On Cognitive Complexity in Discourse. *Journal of Pragmatics*, 40:2003–2026, 2008.

Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, 2004.

Manfred Stede. *Discourse Processing*. Morgan & Claypool Publishers, San Rafael, California, 2011.

Manfred Stede and Carla Umbach. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1238–1242, Montreal, Canada, 1998.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. SRILM at Sixteen: Update and Outlook. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii, 2011.

Aleš Tamchyna and Ondřej Bojar. No Free Lunch in Factored Phrase-Based Machine Translation. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Samos, Greece, 2013.

Sonja Tirkkonen-Condit. Unique Items – Over- and Under-represented in Translated Language? In A. Mauranen and P. Kujamäki, editors, *Translation Universals – Do they exist?*, pages 177–186. John Benjamins, Amsterdam/Philadelphia, 2002.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 252–259, Edmonton, CA, 2003.

F. Ture, D. Oard, and P. Resnik. Encouraging Consistent Translation Choices. In *Proceedings of Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 417–426, Montreal, Canada, 2012.

Bernard Vauquois. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. In *IFIP Congress 68*, pages 254–260, Edinburgh, UK, 1968.

Marc Verhagen and James Pustejovsky. Temporal Processing with the TARSQI Toolkit. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Companion volume: Demonstrations*, pages 189–192, Manchester, UK, 2008.

Marc Verhagen, Inderjeet Mani, Roser Sauri, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, John Phillips, and James Pustejovsky. Automating Temporal Annotation with TARSQI. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Demo Session*, pages 81–84, Ann Arbor, USA, 2005.

Jean Véronis and Philippe Langlais. Evaluation of Parallel Text Alignment Systems: The Arcade Project. In *Parallel Text Processing*, Speech and Language Technology Series, pages 369–388. Kluwer Academic Publishers, 2000.

Yannick Versley. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu, Estland, 2010.

Yannick Versley. Towards Finer-grained Tagging of Discourse Connectives. In *Proceedings of the Workshop 'Beyond Semantics': Corpus-based investigations of pragmatic and discourse phenomena*, pages 145–155, Goettingen, Germany, 2011.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Statistical Machine Translation Output. In *Proceedings of the Fith International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

R. Voigt and D. Jurafsky. Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montreal, Canada, 2012.

Rui Wang, Petya Osenova, and Kiril Simov. Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 119–128, Avignon, FR, 2012.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. Online Large-Margin Training for Statistical Machine Translation. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 764–773, Prague, Czech Republic, 2007.

Ben Wellner. *Sequence Models and Ranking Methods for Discourse Parsing*. PhD Thesis, Brandeis University, Waltham, MA, 2009.

Ben Wellner and James Pustejovsky. Automatically Identifying the Arguments of Discourse Connectives. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 92–101, Prague, Czech Republic, 2007.

Ben Wellner, James Pustejovsky, Catherine Havasi, Roser Sauri, and Anna Rumshisky. Classification of Discourse Coherence Relations: An Exploratory Study using Multiple Knowledge Sources. In *Proceedings of 7th SIGdial Meeting on Discourse and Dialog*, pages 117–125, Sydney, Australia, 2006.

Yorick Wilks. *Machine Translation – It's Scope and Limits*. Springer Verlag, Berlin, Germany, 2009.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conferences on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354, Vancouver, Canada, 2005.

Florian Wolf and Edward Gibson. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics*, 31(2):249–288, 2005.

T. Xiao, J. Zhu, S. Yao, and H. Zhang. Document-level Consistency Verification in Machine Translation. In *Proceedings of MT Summit XIII*, pages 19–23, Xiamen, China, 2011.

Yao Xuchen, Irina Borisova, and Mehwish Alam. PDTB XML: The XMLization of the Penn Discourse Treebank 2.0. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2022–2027, Valletta, Malta, 2010.

Yang Ye, Karl-Michael Schneider, and Steven Abney. Aspect Marker Generation for English-to-Chinese Machine Translation. In *Proceedings of MT Summit XI*, pages 521–527, Copenhagen, Danmark, 2007.

Mohammed J. Zaki and Wagner Jr. Meira. *Fundamentals of Data Mining Algorithms*. Cambridge University Press, Cambridge UK, 2010.

D. Zeyrek, I. Demirsahin, A. Sevdik-Calli, H.O. Balaban, I. Yalcinkaya, and U.D. Turan. The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations. In *Proceedings of the fourth ACL Linguistic Annotation Workshop*, pages 282–289, Uppsala, Sweden, 2010.

Ying Zhang and Stefan Vogel. Significance Tests of Automatic Machine Translation Evaluation Metrics. *Machine Translation*, 24(1):51–65, 2010.

Bowen Zhou, Bing Xiang, Xiaodan Zhu, and Yuqing Gao. Prior Derivation Models for Formally Syntax-based Translation Using Linguistically Syntactic Parsing and Tree Kernels. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 19–27, Columbus, OH, 2008.

Yuping Zhou and Nianwen Xue. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (ACL)*, Jeju Island, Korea, 2012.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, and Jian Su. The Effects of Discourse Connectives Prediction on Implicit Discourse Relation Recognition. In *Proceedings of the 11th SIGdial Meeting on Discourse and Dialog*, pages 139–146, Tokyo, Japan, 2010.

Sárka Zikánová, Lucie Mladová, Jiří Mírovský, and Pavlina Jínová. Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2002–2006, Valletta, Malta, 2010.

Andreas Zollmann, Ashish Venugopal, Stephan Vogel, and Alex Waibel. The CMU-UKA Syntax Augmented Machine Translation System for the IWSLT-06. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT)*, Kyoto, Japan, 2006.

Thomas Meyer

Updated: 31.12.2014

Reherstrasse 21
9016 St. Gallen
Switzerland

Nationality: Swiss
Birth date: April 9th, 1981

+41 77 415 43 09
ithurtstom@gmail.com

Education

2010–2014	**Doctor of Philosophy, Doctoral School of Electrical Engineering (EDEE), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland** Thesis: *Discourse-level features for statistical machine translation* Exams: 4.5/6
2001–2007	Master of Arts UZH, German Linguistics and Computational Linguistics, University of Zurich, Switzerland Thesis: *Kunstvoll-künstliche Textkonstrukte: Zum Unheimlichen der Erzähltechnik E.T.A. Hoffmanns in den "Nachtstücken"*, 5.5/6 Exams: 5.4/6
1996–2001	Matura, in Economics, Cantonal School Burggraben, St. Gallen, Switzerland

Professional Experience

2014–today	**Analytical Linguist, NLP (100%), Google, Zurich, Switzerland**
2010–2014	Research Assistant (100%), Idiap Research Institute, Martigny, Switzerland
2013	Research Visit (100%), Institute for Language, Cognition and Computing, University of Edinburgh, UK
2007–2010	Translation Management (100%), Technical Documentation Department, Metrohm AG, Herisau, Switzerland
2005	Tutor (20%), Lecture on Lexical Functional Grammar (LFG), Institute of Computational Linguistics, University of Zurich, Switzerland
2002–2007	Database Management (30%), Editorial Office M&A Review, University of St. Gallen, Switzerland
2002–2003	Marketing and Opinion Research (30%), Link Institute, Zurich, Switzerland

Research Interests

Computational Linguistics, (Statistical) Machine Translation, Discourse, automated translation of discourse units, Word Sense Disambiguation, Machine Learning techniques applied to Natural Language Processing, Lexical Functional Grammar (LFG) and other formal grammar theories, Corpus Linguistics

Publications

Please see `http://scholar.google.ch/citations?user=OSJzplwAAAAJ&hl=en` or below.

Reviewing

2014	LREC
	ACL Student Research Workshop
2013	ACL Workshop on Statistical Machine Translation (WMT), MT system evaluation
	ACL Workshop on Discourse in Machine Translation (DiscoMT)
2012	KONVENS

Languages

German	Mother tongue
English	Fluent (Cambridge Certificate in Advanced English (CAE))
French	Fluent (3.5 years professional working experience)
Spanish	Reading

Computer Skills

SMT	Moses, Docent, Phrasal
NLProc	Stanford NLP software, LibSVM, Weka, etc.
Programming	Perl, Python, Java, C++, Prolog
CAT, CMS	across, Cosima
Mark-up	XML, XSLT, XQuery, HTML, LaTeX
OS	Ubuntu, Xubuntu, MacOS, Windows

Software and Resources

Code	`https://github.com/ithurtstom`
Annotation	`https://www.idiap.ch/dataset/Disco-Annotation`
Annotation	`https://www.idiap.ch/dataset/Tense-Annotation`

Course Work

EPFL	Machine Learning, Computational Linguistics, Conducting User Studies, Human Language Technology
UZH	Introduction to Computational Linguistics I and II, Programming Techniques in Computational Linguistics I and II, Building Lexicons and Morphology Analysis, Formal Grammars in Linguistics, Selected Techniques of Machine Translation, Formal Grammar and Syntax Analysis, Semantic Analysis, Discourse Analysis, Introduction to Lexical Formal Grammar (LFG), Machine and Machine-Aided Translation, Lexical Resources in Computational Linguistics

Civil Service

2003–2009	Kinderdorf Pestalozzi, Trogen, Switzerland
2003–2004	Stiftsbibliothek, St. Gallen, Switzerland

Personal Interests

Rock climbing, Literature, Socialising, DJ-ing, Computers, Movies, Theatre

Publications

Journals

2014	Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Transactions of Audio, Speech, and Language Processing*, submitted, 2014
2013	Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation-spotting Technique. *Dialogue & Discourse*, 4(2):65–86, 2013
	Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Using the Europarl Corpus for Cross-linguistic Research. *Belgian Journal of Linguistics*, 27:23–42, 2013

In peer-reviewed conferences with proceedings

2014	Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014
	Cristina Grisot and Thomas Meyer. Cross-linguistic Annotation of Narrativity for English/French Verb Tense Disambiguation. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014
2013	Thomas Meyer and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 19–26, Sofia, Bulgaria, 2013
	Thomas Meyer, Cristina Grisot, and Andrei Popescu-Belis. Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 33–42, Sofia, Bulgaria, 2013
	Thomas Meyer and Lucie Poláková. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 43–50, Sofia, Bulgaria, 2013

2012

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA, 2012

Thomas Meyer and Andrei Popescu-Belis. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 129–138, Avignon, FR, 2012

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012

Bruno Cartoni and Thomas Meyer. Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012

2011

Thomas Meyer. Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation. In *Proceedings of ACL-HLT 2011 (49th Annual Meeting of the ACL: Human Language Technologies), Student Session*, pages 46–51, Portland, OR, 2011

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, pages 194–203, Portland, OR, 2011

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR, 2011

Book

2012

Thomas Meyer. *Das Grauen im konstruierten Erzähltext: Zu E.T.A Hoffmanns 'Nachtstücken'*. Diplomica Verlag, Hamburg, Germany, 2012

In reviewed conferences without proceedings

2011

Thomas Meyer, Andrei Popescu-Belis, Jeevanthi Liyanapathirana, and Bruno Cartoni. A Corpus-based Contrastive Analysis for Defining Minimal Semantics of Inter-sentential Dependencies for Machine Translation. In *GSCL-2011 Workshop "Contrastive Linguistics – Translation Studies – Machine Translation – What can we learn from each other?"*, Hamburg, Germany, 2011

Thomas Meyer, Charlotte Roze, Bruno Cartoni, Laurence Danlos, Sandrine Zufferey, and Andrei Popescu-Belis. Disambiguating Discourse Connectives Using Parallel Corpora: Senses vs. Translations. In *Proceedings of Corpus Linguistics Conference*, Birmhingham, UK, 2011

Bruno Cartoni and Thomas Meyer. Building 'Directional Corpora' for Unbiased Contrastive Analysis. In *Proceedings of Corpus Linguistics Conference*, Birmingham, UK, 2011

Technical reports

2012	Nikolaos Pappas and Thomas Meyer. A Survey on Language Modeling using Neural Networks. Research Report 32-2012, Idiap Research Institute, 2012
	Thomas Meyer. Translation Error Spotting from a User's Point of View. Research Report 31-2012, Idiap Research Institute, 2012

Invited talks and presentations

2014	Handling Verb Tense for Statistical Machien Translation, StatMT group, University of Edinburgh, UK, 22 January 2014.
	Discourse-level Features for Statistical Machine Translation. IMS, University of Stuttgart, Germany, 09 January 2014.
2013	Discourse-level Analyses for Statistical Machine Translation. CL colloquium, University of Potsdam, Germany, 24 June 2013.
	Multilingual Annotation of Discourse Connectives: Implications and Applications. Workshop on Discourse Annotation, University of Utrecht, NL, 21 June 2013.
	Verb Tense Labeling and Implicitation of Discourse Connectives in (SM)T. StatMT Group, University of Edinburgh, UK, 22 May 2013.
	Discourse-level Analyses for Statistical Machine Translation. StatMT Group, University of Edinburgh, UK, 27 February 2013.
2012	Automatic Disambiguation of Discourse Connectives. Muldico Workshop II: Multilingual databases and corpora of connectives. University of Jena, Germany, 18-20 October 2012.
	Corpus-based Cross-Linguistic Studies for Machine Translation. Workshop Multilingual Corpora in Cross-linguistic Research, University of Berne, CH, 22 June 2012.
	Using Sense-labeled Discourse Connectives for Statistical Machine Translation. CUSO doctoral winter school, Block 2, Interfaces Syntaxe-Sémantique-Pragmatique, Champéry, CH, 23-26 January 2012.
2011	The Example of the Penn Discourse Treebank. Muldico Workshop I: Towards a multilingual database of connectives, Les Diablerets, CH, 31 August - 02 September 2011.
	Das COMTIS-Projekt. University of Zurich, CH, 19 April 2011.
	Disambiguating discourse connectives for machine translation. CUSO doctoral spring school, Block 2, Interfaces Syntaxe-Sémantique-Pragmatique, Champéry, CH, 14-17 March 2011.