Pronominal anaphora and verbal tenses in machine translation

Loaiciga Sanchez, Sharid

UNIVERSITÉ
DE GENÈVE
FACULTÉ DES LETTRES
Département de linguistique

Doctoral Thesis

Pronominal Anaphora and Verbal Tenses in Machine Translation

Sharid Loáiciga

May 5, 2017

Committee:

Prof. Éric Wehrli, thesis supervisor
Prof. Jacques Moeschler, president
Prof. Joakim Nivre, member
Dr. Andrei Popescu-Belis, member
Prof. Jörg Tiedemann, member

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES LETTRES

IMPRIMATUR

DOCTORAT ÈS LETTRES
Linguistique

Thèse de Madame Sharid LOÁICIGA

Intitulée: *Pronominal Anaphora and Verbal Tenses in Machine Translation*

*

La Faculté des lettres, sur le préavis d'une commission composée de Messieurs les professeurs Jacques MOESCHLER, président du jury; Eric WEHRLI, directeur de thèse; Joakim NIVRE (Uppsala University); Dr Andrei POPESCU-BELIS (IDIAP Research Institute, Martigny); Jörg TIEDEMANN (University of Helsinky), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 5 mai 2017

Le Doyen : Jan BLANC

Thèse N° 882

N.B. – La thèse doit porter la déclaration précédente * et remplir les conditions énumérées dans les informations pour la publication d'une thèse.

Nombre d'exemplaires à remettre à la Direction de l'information scientifique : 2

Abstract

Discourse is the study of texts as sequences of sentences which taken together are able to convey meaning. The properties of a text which enable the conveyance of meaning have been defined as coherence and cohesion. Although they are not independent properties, in this thesis, we focus on reference, a mechanism of cohesion. Concretely, we target pronominal anaphora and verbal tenses. These two linguistic phenomena have the particularity that either a pronoun and its antecedent (the token which gives meaning to it), or a verbal tense and its referent, are potentially placed in different sentences. However, most Machine Translation (MT) systems process texts one sentence at the time.

The problem of pronoun translation has two aspects. First, pronouns are grammatically constrained by their antecedents, and the particular constraints depend on the specific language pair of the translation setting. Second, pronouns are particularly susceptible to translation variations. A pronoun in the source language is not necessarily translated as a pronoun of the same category in the target language. Based on large-scale corpora, we show that the English personal pronouns *it* and *they* can be translated into French by content nominal phrases, by a pronoun of a different category, or be completely omitted.

We look at two different approaches for pronoun translation from English into French: rule-based translation with classic anaphora resolution and cross-lingual pronoun prediction without anaphora resolution. Using the Its-2 MT system (Wehrli and Nerima 2009), we have found that in the former approach, the problem of pronoun translation goes beyond the anaphora resolution problem. Only a subset of the pronouns (*il*, *ils*, *elle*, *elles*) are generated correctly based on the agreement features of their nominal antecedents. This solution is judged, therefore, limited.

The second approach defines a fixed number of classes which includes class OTHER to account for all the non-pronominal translations. Working with the Stanford Maximum Entropy package (Manning and Klein 2003), this approach has allowed us to test the

importance of contextual, morphological and syntactical sources of information, in the form of features, for the correct translation of pronouns. We have found that the immediate context is a particularly good predictor for the high frequency classes *il, ce, ils,* and OTHER, while morphological and syntactical features are beneficial for the *on, elles* and *elle* classes. The *ça* and *cela* classes were not clearly determined by any of these features. Contrary to similar and earlier research, we provide with evidence in favour of using deep linguistic knowledge in the form of syntactic relations to improve performance of pronoun prediction.

Our analysis of the role that different types of features play in the prediction of the pronouns studied here has led us to propose a three-way distinction of the function of the pronoun *it*: pleonastic, nominal anaphoric and event anaphoric. We test the feasibility of this distinction using both Maximum Entropy and Recurrent Neural Network classifiers. We found that separating the nominal anaphoric and event anaphoric realisations of *it* is complex, especially if we consider that, after all, event reference is itself a form of anaphoric reference.

Concerning verbal reference, this thesis follows the line of research that considers verbal tenses as anaphoric and therefore related to the discourse level of language. Verbal tenses require a previously established temporal expression as their referent. For instance, past tenses may have a temporal adjunct, such as *minutes before* or *earlier in the day* as their referent. This work focuses on the translation of the English simple past into French. Building on previous small-scale studies, it validates as significant the translation ambiguity of this tense into French as: *passé composé*, *imparfait*, *passé simple* and *présent*.

The anaphoric relationship between a verbal tense and its referent may span beyond the sentence boundaries. In this work, however, verbal reference is treated at the sentence level only, mainly because we work with a small manually annotated corpus of isolated sentences which makes it impossible to consider the context of the previous sentences.

We consider the usefulness of *grammatical tense* and *boundedness* for the translation of the English simple past into French and compare our work with previous research on *narrativity* for the same task. Grammatical tense is a morphological feature expressed in the pairing of different temporal meanings with different verbal forms. Boundedness refers to an aspectual property of the event used in context, it expresses a property of the sentence in which the verb occurs. Narrativity is a pragmatic property which refers

to the temporal relations holding between events. A narrative relation points to the case when the two events are temporally linked, while non-narrative relations point to the case when events are either not temporally linked or they occur simultaneously. From these properties, our MT experiments show that boundedness produced the best results to improve the translation of the English simple past into French, increasing translation quality measures up to +0.9 BLEU points. Tense improves the translation performance up to +0.5 BLEU points, whereas narrativity improves it up to +0.2 BLEU points.

Throughout all the experiments presented in this thesis, we include complex linguistic knowledge in ways useful to create better MT output with existing MT architectures.

Résumé

Le discours est l'étude des textes en tant que séquences de phrases qui, prises ensemble, sont capables de transmettre le sens. Les propriétés d'un texte qui permettent la transmission du sens se regroupent sous les concepts de cohérence et cohésion. Bien que ce ne soient pas des propriétés indépendantes, nous nous concentrons dans cette thèse sur la référence, un mécanisme de cohésion. Concrètement, nous ciblons l'anaphore pronominale et les temps verbaux. Ces deux phénomènes linguistiques ont la particularité que leurs composants, soit un pronom et son antécédent (le mot qui lui donne un sens), soit un temps verbal et son référent, sont potentiellement placés dans des phrases différentes. Cependant, la plupart des systèmes de traduction automatique (TA) traitent les textes une phrase à la fois.

Le problème de la traduction pronominale a deux aspects. Tout d'abord, les pronoms sont grammaticalement contraints par leurs antécédents, et les contraintes particulières dépendent de la paire de langues considérée. Deuxièmement, les pronoms sont particulièrement sensibles aux variations de traduction. Un pronom dans la langue source n'est pas nécessairement traduit comme un pronom de la même catégorie dans la langue cible. Nous fournissons des preuves à partir des corpus à grande échelle pour montrer que les pronoms personnels en anglais *it* et *they* peuvent être traduits en français par des syntagmes nominaux à contenu, par un pronom d'une catégorie différente, ou être complètement omis.

Nous nous penchons sur deux approches différentes pour la traduction des pronoms de l'anglais vers le français : la traduction basée sur des règles avec résolution d'anaphores classique, et la prédiction multilingue de pronoms sans résolution d'anaphore. En utilisant le système de TA Its-2 (Wehrli and Nerima 2009), nous avons constaté que dans la première approche, le problème de la traduction des pronoms va au-delà du problème de résolution de l'anaphore. Seul un sous-ensemble des pronoms (*il*, *ils*, *elle*, *elles*) est

généré correctement sur la base de l'accord entre leurs traits grammaticaux et ceux de leurs antécédents nominaux. Cette approche apporte donc une solution limitée.

La deuxième approche définit un nombre fixe de classes, y compris la classe OTHER, pour tenir compte de toutes les traductions non-pronominales. À l'aide de la boîte à outils Stanford Maximum Entropy (Manning and Klein 2003), cette approche nous a permis de tester l'importance des sources d'information contextuelles, morphologiques et syntaxiques, sous forme de traits, pour la traduction correcte des pronoms. Concrètement, nous avons constaté que le contexte immédiat est un bon prédicteur pour les classes à haute fréquence *il*, *ce*, *ils* et OTHER, tandis que les caractéristiques morphologiques et syntaxiques sont bénéfiques pour les classes *on*, *elles* et *elle*. Les classes *ça* et *cela* ne sont pas clairement déterminées par aucun de ces groupes d'information. Contrairement à des recherches antérieures et similaires, nous apportons des preuves en faveur de l'utilisation des connaissances linguistiques profondes sous forme de relations syntaxiques pour améliorer la performance de la prédiction des pronoms.

Notre analyse du rôle que les différents types de traits jouent dans la prédiction des pronoms étudiés ici nous a amenés à proposer une distinction tripartite de la fonction du pronom anglais *it* : *pléonastique*, *anaphorique nominale* et *anaphorique événementielle*. Nous testons la faisabilité de cette distinction en utilisant des classifieurs de type Maximum Entropy et de type réseaux neuronaux. Nous avons constaté que la séparation des réalisations anaphoriques nominales et des réalisations anaphoriques événementielles est une tâche complexe, surtout si l'on considère qu'après tout, la référence événementielle est, elle aussi, une forme de référence anaphorique.

En ce qui concerne la référence verbale, cette thèse suit la ligne de recherche qui considère les temps verbaux comme anaphoriques et donc liés au niveau de discours. Les temps verbaux exigent une expression temporelle préalablement établie comme leur référent. Par exemple, les temps verbaux du passé peuvent avoir comme référent un adjoint temporel, comme *quelques minutes avant* ou *plus tôt dans la journée*. Ce travail se concentre sur la traduction du prétérit anglais vers le français. S'appuyant sur des études antérieures à petite échelle, il quantifie l'ambiguïté de traduction de ce temps en français comme *passé composé*, *imparfait*, *passé simple* ou *présent*.

La relation anaphorique entre un temps verbal et son référent peut s'étendre au-delà des limites des phrases. Dans ce travail, cependant, la référence verbale n'est traitée qu'au niveau de la phrase, principalement parce que nous travaillons avec un petit corpus

annoté manuellement de phrases isolées, ce qui rend impossible de considérer le contexte des phrases précédentes.

Nous considérons l'utilité du temps verbal grammatical, *tense*, et de la propriété dite *boundedness* pour la traduction du passé simple anglais vers le français. Nous comparons également notre travail à des recherches antérieures sur la *narrativité* pour la même tâche. Le temps verbal grammatical est une caractéristique morphologique exprimée dans l'appariement de différentes significations temporelles avec différentes flexions verbales. *Boundedness* réfère à une propriété aspectuelle du verbe utilisé dans le contexte, à une propriété de la phrase dans laquelle le verbe se produit. La *narrativité* est une propriété pragmatique qui réfère aux rapports temporels entre les événements. Une relation narrative signifie le cas où deux événements sont temporellement liés, tandis que une relation non narratives signale le cas où deux événements ne sont pas liés temporellement ou ils se produisent simultanément. En utilisant ces propriétés, nos expériences montrent que la propriété boundedness produit les meilleurs résultats pour améliorer la traduction du passé simple anglais vers le français, en augmentant la performance de la traduction jusqu'à +0,9 points BLEU. Le temps verbal grammatical améliore la performance de la traduction jusqu'à +0,5 points BLEU, alors que la narrativité l'améliore jusqu'à +0,2 points BLEU.

Tout au long des expériences présentées dans cette thèse, nous incluons des connaissances linguistiques complexes de manière efficace pour créer de meilleurs résultats de traduction avec les architectures de TA existantes.

Acknowledgements

In the following lines, I would like to thank the incredible people I have come to know over the past few years that this thesis has taken to complete.

Working within the LATL group, my research began closely related to Fips. Even though it gradually shifted away from Fips, I thank my advisor Éric Wehrli for supervising my thesis. I thank Jacques Moeschler for accepting the role of president of the jury and I thank Joakim Nivre, Andrei Popescu-Belis and Jörg Tiedemann for accepting being part of it.

I would have probably not come to Geneva if it was not for the influence of Jorge Antonio Leoni de León. I thank him for introducing me to the field of Computational Linguistics and I thank his family for lending me a helping hand back during my first days in Switzerland.

Before moving to CUI in Battelle, the computational linguistics group was in the middle of the linguistics department at Rue de Candolle. Over there, I met the other computational linguists: Asheesh Gulati, Kamel Nebhi, Luka Nerima, Tanja Samardžić, Jean-Philippe Goldman and Alexis Kauffmann. I thank you all for the many lunches, apéros, discussions and enjoyable working company. I also thank Cristina Grisot, Goljihan Kashaeva, Frédérique Berthelot, Richard Zimmermann, Dara Jokilehto, Karoliina Lohiniva, Anamaria Bentea, Genoveva Puskás, Tabea Ihsane, Joanna Blochowiak, Eva Capitao and all the other great people and colleagues in the linguistics department. Although we were never really at the department at the same time, together with Asheesh, I thank Violeta Seretan and Massimo Brero for all the fun times we have had together.

Back in 2013, I spent six months at IDIAP Research Institute in Martigny. I am grateful to Andrei Popescu-Belis for introducing me to the discourse and statistical machine translation community. I am also grateful to Thomas Meyer for his patience when help-

ing me through all my questions on machine learning and for his constant good humour. I also thank Nikolaos Pappas, Majid Yazdani, Maryam Habibi and Phil Garner for putting up with me during my time there.

In the course of my time at CUI, I very much enjoyed discussing scientific and philosophical issues alike with Kristina Gulordava, Nikhil Garg, Majid Yazdani, Meghdad Farahmand, Sohrab Ferdowsi and Sarah Ouwayda. I thank Paola Merlo for always having some advice at hand to share.

This thesis really came to be during my last year of PhD research at Uppsala University. The time in Uppsala made me a better researcher and produced many happy memories. I thank Joakim Nivre and Beáta Megyesi for letting me to occupy a desk among the computational linguistics group terrific members. I am particularly grateful to Christian Hardmeier for keeping me excited about research, especially when results did not turn out the expected way. I am also happy to have worked with Sara Stymne (who is always ready to help), Fabienne Cap, Miryam de Lhoneux, Marie Dubremetz, Ali Basirat, Yan Shao, Aaron Smith, Eva Pettersson, and Gongbo Tang, some of them part of the 'machine learning for dummies' study group. Thanks Miryam for ensuring the visits to Williams and Fabienne for so many sweet treats. Liane Guillou, Prasanth Kolachina and Amir More also visited Uppsala while I was there. They certainly made me enjoy my own visit even more.

The funding for this thesis has been varied through the years. I thank the Faculté de Lettres for the four different times they have awarded me with scholarships. I am grateful to the COMTIS and MODERN[1] projects for funding my time at IDIAP and for all the meetings I was invited to, even when I was technically no longer part of the projects. I am grateful as well to the Swiss National Science Foundation for providing the funding to attend the 2013 LXML machine learning summer school[2] and specially for making possible my visit to Uppsala University[3].

A lot of moral support has also played a role, perhaps the most important one, to finish this project. I thank my family for encouraging me constantly, even when not fully understanding what I was working on, for so long, and so far away from home. At the end, I would like to express my greatest gratitude to Yves, who has become family, for simply being there –all the way.

[1] Supported by the SNSF, sinergia program, projects CRSI22-127510 and CRSII2-147653.
[2] 10SO13-150933
[3] P1GEP1-161877

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The translation of natural languages using computer programs or Machine Translation (MT) is a field of study pursued since the 1950s, and yet, one in which research seeking to produce high quality translations continues. MT inaugurated the Natural Language Processing (NLP) field in a very rudimentary form, when computer programs were not even conceived to process textual data but numbers (Sparck Jones 1994). Very quickly, however, researchers saw the challenges inherent to natural languages and their translation, in particular syntactic and semantic problems which did not have a clear answer then and still do not have one.

Concerns on the inability of MT systems to understand text have always been present during the development of MT. This was expressed by Bar-Hillel's classic example of a correct translation of the word 'pen' in (1), where the less common reading of 'enclosure' is to be preferred to that of 'writing device', from the probably more common sentence 'the pen is in the box'. (Bar-Hillel 1960; Somers 2003; Hutchins 2010). Although cases like (1) were perhaps purposely tricky, nowadays MT systems would still take ambiguous examples like (2), where the pronoun *it* can refer to either 'water' or to the weather that day, and translate them based on some default translation option.

(1) The box is in the pen.

(2) The water is clear but it is cold.

The decision whether *it* refers to 'water' or to the cold temperature of the day depends

on the context. If the sentence is taken in isolation, it is difficult to tell which interpretation should be preferred. However, like for many NLP applications, the sentence is the working unit of practically all MT systems. The sentence as the initial symbol S of a grammar was a formalization of generative grammars in the 1950s as well. The sentence as a unit is indisputably a practical construct. It constitutes a natural boundary very useful for splitting a text so it can be processed *one sentence at the time*, but this type of processing also entails the assumption that sentences are independent from one another.

Although the progress in MT quality since the foundational work is undeniable, the one-sentence-at-the-time processing has not been without its disadvantages. Intersentential dependencies such as pronouns and their referents are an example of a linguistic phenomenon which extends beyond sentence boundaries. An example is presented in (3), where the original English sentence (3a) has been translated by a human (3b) and by an MT system (3c) built for one of our experiments (Section 4.3). As expected, in the human reference translation (3b), the referent *authorities* is translated with the feminine *autorités* and hence the pronoun *they* is translated with the French feminine pronoun *elles*. For an MT system (3c), on the other hand, the relationship is not perceived since the referent is in a sentence processed previously to the sentence containing the pronoun. Therefore, the pronoun is translated with a masculine pronoun as result of statistical learning, since it happens to be more frequent overall than the feminine pronoun. We confirmed the problem using Google Translate[1].

(3) a. SOURCE The Republican *authorities* were quick to extend this practice to other States. Over the past two years, *they* sponsored bills in 34 States to force voters to show a photo ID card.

b. REFERENCE Les *autorités* républicaines s'empressèrent d'étendre cette pratique à d'autres États. Au cours des deux dernières années, *elles* parrainaient des projets de loi dans 34 États pour forcer les électeurs à présenter une carte d'identité avec photo.

c. MT SYSTEM Les *autorités* républicains ont été prompts à étendre cette pratique à d'autres États. Au cours des deux dernières années, *ils* ont parrainé factures dans 34 États à forcer les électeurs de montrer une photo cartes d'identité.

[1] https://translate.google.com/#en/fr/, used on January 30th, 2017.

Examples like (3) suggest that, given the quality of translation reached by sentence translation, text-level translation needs to be considered next. While a variety of linguistic definitions of the term 'discourse' have been suggested, in this thesis *text-level* or *discourse* processing are synonym terms and they are understood as language processing beyond the sentence level (Stede 2012). This definition belongs to the language technology domain and suits our work since it is concerned with MT for written texts and hence written discourse, with a clear beginning and end (Webber, Egg, and Kordoni 2011).

In this sense, discourse studies take texts as (ordered) sequences of sentences able to convey a meaning as a whole and not as just the sum of each of its parts. That a text has an internal order is a reasonable assumption behind sentence-level alignment of parallel corpora for example, where sentence correspondences are found because documents have a similar and systematic structure in both languages. Order alone, however important, is not enough to convey the meaning of a text or discourse, other properties are involved as well. Discourse structure has two properties which are genre-independent and are in constant interaction: *coherence* and *cohesion*. The first is interested in making a text understandable. It concerns the non-explicit logical relationship, i.e., coherence relation, between two text spans (clauses, sentences, paragraphs, etc.). Cohesion is interested on the surface clues that link and hold a text together, giving it connectivity. Example (4a) offers a simple text which exemplifies these two properties. In contrast, (4b) presents a similar text where all the elements are linked together, but there is no logical sense; (4c) presents a text which is somewhat understandable, but where the words linking the sentences together are missing. (4b) has cohesion but not coherence, while (4c) has coherence but not cohesion.[2] Cohesion concerns concrete linguistic devices of *reference*.

(4) a. My favorite colour is blue. I like it because it is calming and it relaxes me. I often go outside in the summer and lie on the grass and look into the clear sky when I am stressed. For this reason, I'd have to say my favorite colour is blue.

b. My favorite colour is *blue*. *Blue* sports cars go *very fast*. *Driving* in this way is dangerous and can cause many *car crashes*. I had a *car accident* once and *broke my leg*. I was very sad because I had to miss a holiday in Europe because of *the injury*.

[2]These examples have been taken from Scrouton (2011).

c. My favorite colour is blue. I'm calm and relaxed. I'd have to say my favorite colour is blue.

Reference is defined as the relationship between a linguistic expression and a world entity or *referent* which gives meaning to it (Halliday and R. Hasan 1976; Zufferey and Moeschler 2012). Research on reference[3] and MT has focused on three types of linguistic devices pointing to three types of reference: connectives (adverbial reference), verbal tenses (temporal reference) and pronouns (nominal reference). While connectives have been largely treated by other researchers for English-French translation, verbal tenses have been mostly treated by researchers interested in the translation from Chinese to English. Pronouns, on the other hand, have been at the center of the anaphora and coreference resolution tasks. The anaphora resolution task links pronouns with their antecedents or referents, while the coreference resolution task links all nominal expressions to the same referent together. The independent development of MT has added a new dimension to these tasks: multilingualism, giving them new perspectives. However, MT is not concerned with finding referents per se, it is rather interested in the problem of how to improve the translation of pronouns.

Coherence and cohesion are both necessary for appropriate text understanding and they do not exist one without the other. In this work, however, we part from the assumption that reference is a property of cohesion and we focus on two facets of reference and its treatment and implications for MT. We are interested in the translation of pronouns and verbal tenses. The human processing of these two phenomena involves several linguistic levels spanning from morphosyntax to pragmatics. For pronouns, the lexical properties of the languages involved, as well as their principles of morphological agreement and syntactic binding interact. For verbal tenses, the interpretation cues for their correct understanding come from several elements as diverse as the lexical choice, the adverbs and the particular type of clauses used.

1.1 Pronominal reference and machine translation

Pronouns are the most prominent kind of nominal reference (Mitkov 2002), i.e., nominal expressions which designate a same entity within a text in a recurrent manner. Pronouns

[3] All references to the works introduced in this chapter will be given in Chapter 2.

are economical, short and independent words which can stand in the place of a more cumbersome word, hence, they lack some capacity to convey lexical meaning (De Beaugrande and Dressler 1981). Why languages are equipped with pronouns is not clear from the literature. De Beaugrande and Dressler (1981, p. 65) argue that their main purpose is to avoid unnecessary repetition of concepts, providing access to semantic content with efficiency and "less processing effort by being shorter than the expressions they replace".

When translating from one language to another, pronouns must be mapped. In particular, the agreement information between antecedent and pronoun must be preserved according to the target language. In machine translated texts, however, the agreement information is often lost, producing a mismatch between the pronoun and its referent and compromising the text's appropriate understanding. Several studies have confirmed this observation. Brennan, Freidman, and Pollard (1987), for instance, express that inappropriate use or failure to use pronouns causes communication to be less fluent. Hardmeier and Federico (2010) state that mistranslation of pronouns renders the MT output hard to understand even when content words are not affected. Guillou (2012), in addition, says that an incorrect pronoun translation can result in a misleading and confusing reading of the text.

The problem of pronoun translation has two axes. First, while pronouns are constrained in all languages by their antecedents, the constraints depend on the specific pair of languages one is dealing with (Webber, Egg, and Kordoni 2011). For instance, the English personal pronoun *it* can have up to 12 translations in French (*il, elle, la, le, lui, cela, celui, celui-ci, celle-là, ce, en, y*) depending on the gender and number of the antecedent and the grammatical function, two of which, in addition, have phonologically constrained variants (*l', c'*) (Popescu-Belis et al. 2012). Second, pronouns are particularly susceptible to translation variations. For instance, while it is relatively safe to assume that a verb is translated by a verb, pronouns are not always translated as pronouns. They can correspond to a content nominal phrase, they can be completely omitted or translated by a pronoun of a different category. This is the case in example (5a) where the English pronoun *it* is translated with the nominal phrase *ce projet* in (5b) and with the adverbial pronoun *y* in (5c). One could argue that one option or the other depends on the human translator who could just as well choose a translation with a pronoun of the same category systematically. However, the human translator takes into consideration the rest of the document and the stylistic effect of the writing. For instance, whenever a contrastive effect is intended, the use of pronouns is discouraged (Garnham 2001). As it will be

shown in Chapter 4, translations like (5b) and (5c) are not at all uncommon, although the translation in (5d) is more literal.

(5) a. SOURCE Financially, PSG has the means to make *it* happen.
b. REFERENCE 1 Financièrement, le PSG se donne les moyens pour que *ce projet* se concrétise.
c. REFERENCE 2 Financièrement, le PSG a les moyens pour *y* parvenir.
d. REFERENCE 3 Financièrement, le PSG a les moyens de *le* réaliser.

For a human, resolving anaphoric relations is more or less straightforward. For an MT system, on the other hand, the relationship is not handled unless there is some explicit linguistic knowledge involved. In this work, our goal is to understand the linguistic factors preventing the correct translation of pronouns. In Chapter 5, we will examine these factors using both a rule-based approach and a classification approach. Classification experiments allow us to isolate each possible variable and understand its role in the generation of translated pronouns. Furthermore, in Chapter 6, we will examine the effect of the function of the pronoun on the translation. In particular, we will look into event reference pronouns, which refer back to verb phrases, predicates or entire clauses.

1.2 Temporal reference and machine translation

The referential structure of a text concerns events – states or actions – and their situation with respect to the moment of speech and with respect to each other (Zufferey and Moeschler 2012). The interpretation clues to assign a temporal value to events are encoded in verbs[4], or more precisely verb phrases and their tense, aspect and mode (TAM) features. These characteristics place the events at a particular time line, encode the perception of the speaker about them, and express their level of factuality respectively (Aarts 2011). The TAM features are known to be encoded quite differently across languages. In a context such as MT, they can cause divergences when matching the translation of verbal tenses, for instance, if translating from a language in which a single form may correspond to several forms in the target language. This scenario is typical when translating into a morphologically richer language from a less rich one. The problem has been mentioned by Vilar et al. (2006) for the translation of English to Spanish. For instance,

[4]This is the case for the major European languages with which this work is concerned.

in Spanish both verbs *ser* and *estar* are translations of the English verb 'to be'. The first one is used for permanent properties of objects or people, and the second one is used for expressing temporary qualities. MT systems, however, do not distinguish between these two verbs. In addition, they note that Spanish has 17 verbal tenses forms for each the indicative and the subjunctive moods, which have no direct correspondence into English. Comparable problems have also been mentioned by Silva (2010) for the translation of Brazilian Portuguese to English.

This thesis follows the line of research which considers verbal tenses referential and therefore anaphoric (Partee 1973; Partee 1984; Moens and Steedman 1988). The referential relationship of verbal tenses is in some respects similar to that of pronouns with an event reference function – which refer back to verb phrases, predicates or entire clauses – in the sense that the referent is not limited to a specific lexical item as in the case of the nominal antecedents of pronouns. Verbal tenses have been argued to find their referent in the adverbials, when-clauses and other contextual clues in the sentence, changing their specific temporal interpretation according to the particular sentence they appear in. Existing MT systems do not take these sources of temporal interpretation into account when translating verbal tenses.

In our work, we explore the usefulness of tense (Chapter 7) and aspect (Chapter 8) as features for correctly disambiguating the English simple past when translating into French. This English tense has four frequently used translations in French: *passé composé*, *imparfait*, *passé simple* and even *présent*. The choice of the verbal tense depends on the fine-grained temporal interpretations of the utterance in context; however, machine translated texts show a skewed distribution in favor of the *passé composé* translation, yielding confusing translations which are not fluent. As an illustration, in example (6) we show a sentence translated into French by a baseline system built for our experiments (Section 8.5). In the example, both verbs in bold are English verbs in the simple past tense translated using the French *passé composé* tense. For the second one, however, the reference translation proposes *imparfait* tense, given the unfinished perspective in time expressed by the verb *were*.

(6) a. SOURCE In defense of his policy he *added*, however, that these wars *were* essential in order to bring about peace, despite the high cost.

b. MT SYSTEM Dans la défense de sa politique, il *a ajouté*, toutefois, que ces guerres *ont été* des éléments essentiels dans le but de favoriser la paix,

malgré le coût élevé.

c. REFERENCE À la décharge de sa politique il *a ajouté* que dans certaines situations la guerre *était* indispensable pour obtenir la paix, même si le prix en est élevé.

The main aim of this thesis is not to improve the existing theory of tense and aspect, but rather to exploit existing theories and resources related to them in order to improve the MT of verbal tenses between English and French. Our interest is to provide a linguistic perspective into the translation process of this particular reference device. Indeed, the existing literature on tense and aspect issues is large but working formalizations for NLP applications are scarce.

The first exploits the syntactic knowledge intrinsic to the parser the translation system is built on to find antecedents for the pronouns. This resulted in accurate translations for the cases in which a pronoun have a nominal antecedent and the system is able to find it. But it also provided evidence to the fact that not all pronouns are translated by a pronoun of the same category, partly because not all of them have a nominal antecedent. Given what we know about the distribution of the translation of pronouns, it is difficult to create enough rules that generalize all the translation possibilities in all possible contexts.

Our cross-lingual pronoun prediction experiments, on the other hand, have allowed us to model different types of information and to test their predictive power over a limited number of classes. Using this approach, we have also provided evidence in favor of including syntactic knowledge for the task. We have found that syntactic information combined with enough contextual information could represent an alternative to the information provided by external anaphora and coreference resolution systems.

1.3 Contributions

The main contribution of this thesis is the application of linguistic insights on two discursive aspects of MT: pronominal anaphora and verbal tenses. Concerning pronoun translation, we have investigated the strengths of contrastive system architectures and the role of diverse types of features. In particular, we propose a three-way distinction of the pronoun *it* based on its function as *anaphoric*, *event* and *pleonastic*, which may benefit not only MT but also the task of coreference resolution. Our treatment of verbal

tenses, on the other hand, has exploited the concepts of grammatical tense and actualization aspect in real-scale SMT settings. The translation of verbal tenses has only been addressed lightly within the MT domain. We have provided two successful case studies of formalizing deep linguistic knowledge and robustly passing it to SMT systems, resulting in improvements of translation.

In the next paragraphs, the major contributions of this work are described, following the order in which they appear in the thesis.

- In Chapter 4, we first present extensive corpus analyses on the distribution of pronouns in different genres, corpora and languages. We show that the translation of each pronoun presents a multiple choice depending on the particular preferences of the target language.
- In Chapter 5, we have evaluated a rule-based system and its integrated anaphora resolution procedure within the context of a shared task. Currently, rule-based systems are developed less and their comparison with other architectures is seldom done.
- Moreover, in Chapter 5, we describe our study of different types of linguistic features for the cross-lingual pronoun prediction task. Concretely, we compare the usefulness of syntactic, morphological and contextual features and provide evidence in favor of including syntactic knowledge.
- Working with the ParCor corpus (Guillou, Hardmeier, Smith, et al. 2014), in Chapter 6, we assess the feasibility of the three-way disambiguation of pronoun *it* based on its function in text as nominal anaphoric, event anaphoric or pleonastic. We present systems trained on both gold-standard data and silver-standard data.
- Furthermore, we have contributed an annotated parallel corpus with English and French tenses. The annotation process is automatic and it is explained in Chapter 7.[5] The set of rules used to compute the English and French tense labels has prompted the development of a tool to annotate raw data with tense labels. The tool enhances and increases the rules presented in Appendix A and additionally provides rules for German. The system description paper is accepted for publication and will appear soon (Ramm et al. 2017).
- Finally, in Chapter 8, we have used the annotated corpus provided by Grisot and

[5]The corpus can be downloaded at `https://www.idiap.ch/dataset/tense-annotation`

Cartoni (2012) which contains information on different aspects of temporal reference to annotate new unlabeled data to train a phrased-based statistical MT system.

– Our work in Chapters 6 and 8 provides concrete examples of methods for utilizing existing and limited linguistic resources to successfully achieve large-scale MT experiments.

1.4 Relation to published work

The work presented in this thesis expands and elaborates on published work. Most of these publications are the result of collaborative efforts with different colleagues over the years. Below, we describe my participation in the publications with multiple authors.

– I did part of the French and English manual evaluations for the corpus study presented in Scherrer et al. (2011). The analysis presented in this thesis does not appear in the paper.

– I worked on the rules for the anaphora resolution module of the Its-2 rule-based translation system presented in Loáiciga and Wehrli (2015). In addition, I completed all the automatic and manual evaluations described in the papers and presented here. Yet, this MT system, along with the Fips parser it is based on, has been developed exclusively by Éric Wehrli and Luka Nerima (Wehrli 2007; Wehrli and Nerima 2009) for many years.

– For the tense experiments (Loáiciga, Meyer, and Popescu-Belis 2014), I designed and completed the automatic annotation and I trained all the SMT systems. The evaluations were done equally by all the authors, like the writing of the paper. The classifiers for the prediction of tense presented in the paper were implemented by Thomas Meyer.

– Concerning the work on aspect and MT published in Loáiciga and Grisot (2016), I outlined and completed all the classification and MT experiments, including the feature engineering and significance tests. Cristina Grisot was responsible for the description of the gold corpus while I wrote the other parts of the paper. The manual evaluation of the classifier labels was done by Cristina Grisot, while the manual evaluation of the MT output was done by both authors.

– For the work on disambiguation of *it* presented in Loáiciga, Guillou, and Hard-

meier (2016), Liane Guillou provided the data extraction and description. Feature engineering was worked out by all three of us. I implemented the classifiers and did all the manual evaluations. The implementation and description of the source-aware language model was completed by Christian Hardmeier.

– Last, the work on disambiguation of *it* led to another paper currently under review (Loáiciga, Guillou, and Hardmeier SUBMITTED). For this paper, I completed all the experiments with the supervision of Christian Hardmeier. The manual error analysis was carried out by Liane Guillou and myself. I wrote most of the paper.

1.5 Conclusion

Both coherence and cohesion are linguistic properties of a discourse structure. They hold together pieces of information, making a text comprehensible, and not just a group of sentences put together. Reference, a cohesive property, provides with a mechanism of continuity which assists the logical conveyance of meaning or coherence.

The goal of this dissertation is to investigate two different devices of discursive reference, i.e., pronouns and verbal tenses and their implications for MT. Our aim is two-fold. One the one hand, we want to understand the linguistic factors determining the referential elements of a correct pronoun translation and a correct verbal tense translation. We will achieve this through classifiers which allow to control variables in the form of features. On the other hand, our experiments will use this knowledge in conjunction with different MT system architectures, in particular, rule-based, phrase-based and factored models of translation. We expect to better understand the ways in which linguistic knowledge can leverage the MT process. From a linguistic perspective, studying these two aspects cross-linguistically and using large corpora and computational techniques brings new insights. From a language technology perspective, the linguistic knowledge contributes to create more fluent and natural translations.

Chapter 2

Related research

Pronominal anaphora and verbal tenses are associated with the discourse level of language. Much of the earlier work interested in NLP and discourse, in fact, addresses *discourse structure* segmentation. More recently, and for MT in particular, much more work exists explicitly interested in one aspect of discourse, such as lexical consistency, pronominal reference, tense and aspect and connetives. This last one in particular has been largely addressed in recent studies, due to the relevance of connectives for discourse structure segmentation and identification.

In this work, we investigate both pronouns and verbal tenses as mechanisms of pronominal and temporal reference respectively. Research on pronominal reference has traditionally been the object of study of the anaphora and coreference resolution domains (AR). Besides some efforts to integrate an AR strategy in early rule-based MT, it is only from 2010 onwards and within the statistical MT paradigm that pronoun translation placed itself as discourse-related research. In our own experiments, we will use both rule-based and statistics-based approaches to pronoun translation. The problem of temporal reference, on the other hand, has been tackled mainly in the context of the translation from and into Chinese, due to the wide gap between tense prominent languages such as English and aspect prominent languages such as Chinese. Our work represents most of the existing efforts to treat the translation between English and French.

2.1 Discourse in NLP

As explained by Webber, Egg, and Kordoni (2011), a text or discourse forms a comprehensible text when the patterns formed by its sentences convey more meaning than each sentence alone. For this transmission of meaning, a text exploits several language features. These features relate to the *coherence* and *cohesion* of a text. They ensure that the text forms a single whole, in the case of coherence ensuring that each utterance is an appropriate sequel of a preceding utterance in a logical manner. In the case of cohesion, several linguistic devices achieve a continuity of reference to the same objects: referential expressions, connectives, verbal tenses, ellipses, word repetitions, related words (Halliday and R. Hasan 1976; Zufferey and Moeschler 2012; Stede 2012).

Researchers working in the language technology domain noticed problems related to the text or discourse level very early, for example in question-answering systems which could not answer successive questions with pronouns (Webber, Egg, and Kordoni 2011). These initial problems where addressed using heuristics, but it soon became clear that a lot of implicit information in texts needed to be inferred and therefore modeled. Eventually, from the semantics and pragmatics theoretical approaches to discourse, some theories such as Centering Theory (Grosz, Joshi, and Weinstein 1986; Grosz, Joshi, and Weinstein 1995), Rhetorical Structure Theory (RST) (Mann and Thompson 1988) and Segmented Discourse Representation Theory (SDRT) (Lascarides, Asher, and Oberlander 1992; Lascarides and Asher 1993) were developed. Webber and Joshi (2012) provides a survey of these works. This line of research focused on developing an underlying structure of discourse. Congruently with advances in syntax theory which proposed an underlying structure of the sentence (Chomsky 1972; Chomsky 1981), the first type of discourse structures proposed were trees. More recently, directed acyclic graphs structures and linear structures have been produced (Webber, Egg, and Kordoni 2011).

2.1.1 Discourse structure

Finding the discursive structure of a text involves segmenting it in units or text spans and then labelling those units following semantic and pragmatic principles. A *coherence relation* refers to the often non-explicit relationship between two discursive units. Early discursive theories suggested different rules for achieving that process while attempting to formalize the discursive property of coherence. This means that individual

units of information are expected to be meaningfully related to one another with a sense between adjacent units, giving a text an inner logic built up by a particular order of the units that compose it and by a common topic (Webber, Egg, and Kordoni 2011; Stede 2012; Webber and Joshi 2012). While the main goal of discourse parsing is building the discourse parse structure of a text, discourse structure raises some constraints for anaphora resolution.

Rhetorical Structure Theory (RST)

RST defines a coherent relation in terms of the intentions of the speaker or the writer. Most relations are said to hold between a unit that is more important or the *nucleus* and one unit that is ancillary, or supportive in nature, called the *satellite*. 25 different main relations have been defined and are meant to apply to all texts. RST is claimed to be empirical since the relations have been collected from different texts (letters, advertisements, scientific articles, newspaper, etc.), but they remain subjective intuitions of the researcher (Mann and Thompson 1987; Mann and Thompson 1988).

RST can be grasped better through the example in (1), taken from Mann and Thompson (1988). In this example, each Elementary Discourse Unit (EDU) is delimited with square brackets and identified with a number. An EDU is a unit of information or text span denoting an event of type of event, which does not always match the sentence boundaries. Discourse parsing basically consists in EDU segmentation and assigning coherence relations between EDUs for building Elementary Discourse Trees (EDT) as the one presented in Figure 2.1.

(1) a. [Concern that this material is harmful to heath or the environment may be misplaced.]$_1$

b. [Although it is toxic to certain animals,]$_2$ [evidence is lacking that it has any serious long-term effect on human beings.]$_3$

The automatic segmentation of discourse into a sequence of EDUs and the identification of the relations between these EDUs is appealing for several NLP applications.

In SMT, for instance, Ghorbel, Ballim, and Coray (2001) proposed that EDUs are useful for parallel corpus alignment based on semantic content. Their particular work concerns the alignment of medieval texts with their modern translations, but their general idea is

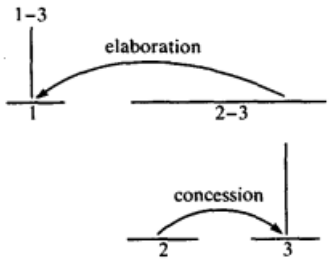


Figure 2.1: Elementary Discourse Tree representing example (1)

that text level alignment of texts where the punctuation, paragraphs and sentences are structured very differently can be helped by discourse structure segmentation.

Moreover, the translation process itself has been claimed to benefit from discourse structure knowledge. Anchored on the discrepancies between the discourse structures of Japanese and English, Marcu, Carlson, and Watanabe (2000) argue that a MT system able to identify the mismatches between the EDUs and discourse relations of the source and target languages would yield more natural and fluent output.

Recently, RST has also been used as the basis for a new MT evaluation metric called DiscoTK (Joty et al. 2014). This measure builds on the output of an RST discourse parse tree of the hypothesis and the reference translations, and evaluates the matching coherence relations between the two. This metric obtained the best correlation scores with human judgments in the shared evaluation task of the Ninth Workshop on SMT (WMT 2014).

Segmented Discourse Representation Theory (SDRT)

While RST is largely used for discourse parsing, SDRT is also a known discourse structure theory. In SDRT relations are motivated by syntactic and semantic principles. They can be coordinating (of the same level) or subordinating (parenthetical) and are limited to NARRATION, ELABORATION, EXPLANATION, BACKGROUND, EVIDENCE, CONSEQUENCE and CONTRAST (Asher and Lascarides 1995).

How to apply these discourse relations in order to segment the discourse content is determined by a formal logic mechanism which provides with four types of rationale to achieve inferences (Lascarides, Asher, and Oberlander 1992; Lascarides and Asher

1993; Lascarides and Asher 2007). Probably because these formal-logic principles are difficult to implement, there is not much empirical work based on SDRT. Examples of implementations include Yllescas (2012)'s discourse parsing system for Spanish and Asher, Denis, et al. (2004)'s system for anaphora resolution.

One very important contribution of this theory is the *right frontier constraint*, first postulated by Polanyi (1988) and further advanced by Asher (1993) and Asher (2005). This constraint states that every new discourse constituent must be attached on the right frontier of the ongoing discourse tree. This has a critical implication for anaphora interpretation, since potential antecedents are limited to those on the right side of the discourse structure tree.[1] The constraint was put to test with participants in an experimental framework by Holler and Irmen (2007). Participants were presented with short passages of six clauses with an anaphor in the last line and two potential antecedents, one in the first and one in the fourth line, as presented in (2a) - (2b). To chose the correct antecedent, it was found that gender is the first disambiguation criterion, but when potential antecedents have the same gender, then the placement to the right (right frontier constraint) is the disambiguation criterion (Fig. 2.2).[2]

(2) a. $_{\pi 1}$In the morning *the student* went to the university $_{\pi 2}$because it was time to attend the lecture on advantages and disadvantages of Kant's categorical imperative. $_{\pi 3}$The lecture hall was busy. $_{\pi 4}$*The fellow student* was as always in a bad mood $_{\pi 5}$because nobody listened.

b. In the afternoon *she* still had many things to do.

2.1.2 Reference

Discourse structure is rather linked to the property of coherence. Indeed, a discourse tree is composed by the logical association existing between two text segments and the

[1] This is one of the main innovations compared with the Discourse Representation Theory (DRT) (Kamp and Reyle 1993; Eijck and Kamp 1997), which the SDRT is built upon. The motivation of DRT was to interpret nominal and temporal anaphora in discourse; however, its principles were not enough to disambiguate anaphors.

[2] Note that the original experiment is in German and the two potential antecedents are therefore inflected for gender.

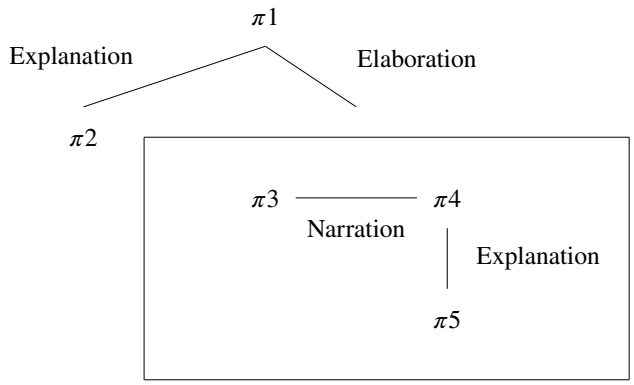


Figure 2.2: SDRT discourse tree for the example (2a). The preferred interpretation of the pronoun in (2b) is as coreferential with the antecedent in $\pi 4$ due its placement to the right in the tree.

different coherence relations between them. There are, however, other important factors ensuring the comprehension of a text. The segments of a text are also linked by means of *ties* or mechanisms of *reference* to the same entities. Reference evokes the property of cohesion of a text, and although it is not completely dissociated from coherence, it is more about concrete elements of interpretation, "where the interpretation of any item in the discourse requires making reference to some other item in the discourse, there is cohesion" (Halliday and R. Hasan 1976, p. 11).

There are different types of reference depending on the type of linguistic expression involved. Among them, nominal reference is perhaps the most prominent one since it includes pronouns which are extremely frequent and necessarily invoke another referring expression to express lexical meaning. Other major categories of reference include comparative reference (involving some keywords such as *same, equal, identical, identically, such, similar, so similarly, likewise, other, different, else*, etc.), temporal reference or reference to events (through verbal tenses), and adverbial reference (involving anaphoric discourse adverbials such as *however, because, in the meantime, etc.*), (Mitkov 2002; Mitkov 2003; Webber, Stone, et al. 2003).

Concerning MT, research on reference revolves around three types of linguistic expressions pointing to three types of reference: connectives (adverbial reference), pronouns (nominal reference) and verb tenses (temporal reference). Adverbial reference, has been largely treated by other researches, in particular in the work by Popescu-Belis et al. (2012), Meyer, Popescu-Belis, Zufferey, et al. (2011), Meyer and Popescu-Belis (2012),

Meyer, Popescu-Belis, Hajlaoui, et al. (2012), and Meyer (2014). In our work, we are interested in the last two types of reference, pronominal and temporal. The following two sections examine the relevant research on each one, including both theoretical and computational approaches.

2.2 Pronominal reference

Pronouns are referential expressions devoid of lexical content, they need context which provides them with a referent or *antecedent* to find meaning. Pronouns are a type of referring expressions, i.e., expressions linked to a world entity. Indefinite noun phrases, definite nouns phrases, proper names, and demonstratives constitute different types of referring expressions as well (Zufferey and Moeschler 2012). The term *coreference* is used to describe the fact that different referring expressions, known as *mentions* in computational approaches, may point to the same world entity, forming a coreference chain. *Coreference resolution* (CR) refers to finding all mentions in a text and classifying them into *chains*, each chain corresponds to a referent or world mention. *Anaphora resolution* (AR), on the other hand, refers to the relatively more specific task of finding the antecedent for each anaphor or pronoun in the text, i.e., the expression which gives meaning to it.

2.2.1 Early approaches to pronominal reference

Binding Theory

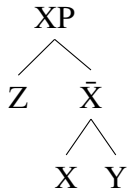
Within syntactic theory, referring expressions are analyzed using Chomsky's *Binding Theory* (Chomsky 1980; Chomsky 1981; Büring 2005), a mechanism that rules and restricts how referring expressions corefer within a sentence. BT does not point to a referent directly, but allows to discriminate between potential antecedents for a pronoun within a sentence. While binding theory distinguishes between the three types of referring expressions listed in (3), only (3a) are considered *anaphors*. (3c) includes non-pronominal noun phrases such as proper names and are known as *r-expressions*.

(3) a. reflexives and reciprocals
b. non-reflexive pronouns

c. r-expressions

Since binding theory is part of Chomsky's *government and binding* grammar model, it is based on the notion of *c-command*. This is a hierarchical representation which provides an explicit formulation of the grammatical constraints to interpret nominal phrases and pronouns; consequently, these constraints can be used to discriminate a potential antecedent of a pronoun according to its position. In any given tree, a node A c-commands a node B iff 1) A does not dominate B and B does not dominate A and 2) the first branching node dominating A also dominates B. As expressed by Haegeman (1994), "node A will be said to dominate node B if you can go from node A to node B along a downward branch". Example (4) shows how X and Y fulfill the two conditions: i) neither of them dominate each other and ii) the first branching node dominating X, i.e., $\bar{X}$, also dominates Y. Therefore, X and Y c-command each other. Furthermore, the A and B are said to be *bound* if conditions in (5) hold.

(4)



(5) a. A c-commands B
b. A and B are coindexed (they agree)

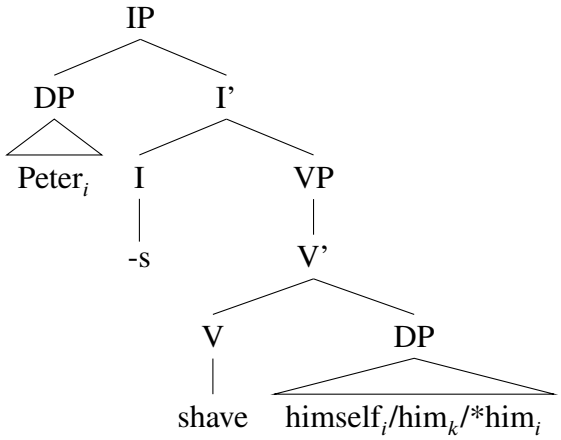
Each of the three types of referring expressions involves a principle based on the notions of *bound* (i.e., contained) and *governing category* (i.e., the smallest clause that includes them). They are listed in (6).

(6) a. Principle A: reflexives and reciprocals must be bound within their governing category.
b. Principle B: other pronouns must be free (not bound) in their governing category.
c. Principle C: r-expressions must be free.

To understand these principles better, take for instance sentences in (7). The pronoun *himself* in sentence (7a) is grammatical because it is coindexed with *Peter*, i.e. they agree

in all their features, and because *Peter* c-commands *himself*, i.e. the first branching node dominating *Peter* also dominates *himself*. Therefore, according to conditions (5a), (5b), they are bound, and according to the **principle A**, we can deduce that the only antecedent to *himself* is *Peter*.

(7)



a. Peter$_i$ shaves himself$_i$.
b. Peter$_i$ shaves him$_k$.
c. *Peter$_i$ shaves him$_i$.

Following this same reasoning, **principle B** validates (7b) and rules out (7c); since *him* is a pronoun it cannot be bound, i.e., find its antecedent within the same sentence. Note that *bound* is interpreted as within the sentence level, which is the basis of the theory itself.[3]

Centering Theory

Another important theory for pronoun interpretation is *Centering Theory* (Grosz, Joshi, and Weinstein 1986; Grosz, Joshi, and Weinstein 1995). Similarly to binding theory, it consists in a set of rules and constraints that govern the relationship between what the discourse is about and some of the linguistic choices of syntactic structures and referring expressions. It states that at any point in a discourse (i.e., a text or a conversation), there is one single entity that is the most salient discourse referent at the point. This referent,

[3]Note as well that according to the BT the term *pronoun* has a very strict definition, it includes reflexive and reciprocal pronouns exclusively.

the current focus of attention, is called the *center*, and the main goal of the theory is how to compute the centers. Concerning pronouns, "there is a pronominalization rule that says whenever some discourse referent is referred to by a pronoun, then in the same sentence the center referent must *also* be pronominalized" (Stede 2012, pp. 51-52). This rule is illustrated in example (8), taken from Grosz, Joshi, and Weinstein (1995). In this example, since *Terry* is the center or the subject, one needs to repeat *Tony* in (8e), because only *Terry* can be pronominalized.

(8) a. Terry really goofs sometimes.
b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.
c. He wanted Tony to join him on a sailing expedition.
d. He called him at 6 AM.
e. ?He/Tony was sick and furious at being woken up so early.
f. He told Terry to get lost and hung up.
g. Of course, ?he/Terry hadn't intended to upset Tony.

(9) In the group there was one person. It was Mary who left.

Both the binding theory and the centering theory account for third-person personal pronouns and a well defined type of reference to an antecedent; however, other types of nominal reference exist. These include deictics or demonstrative pronouns used to refer to one or more clauses (Webber 1990). These also include bridging, a referential relationship between two elements linked through inference and not necessarily morpho-syntactical information; for instance, the referential link between *one person* and *Mary* in example (9) (Clark 1975, p.170). Finding the referent or antecedent for these types of nominal reference is considered as a highly difficult task (Poesio and Vieira 1998; Ferrández and Peral 2000; Mitkov 2002; Mitkov 2003).

2.2.2 Development of anaphora and coreference resolution systems

Foundational systems

The Anaphora Resolution (AR) task has occupied theoretical and computational linguistics since the late 1970s. Most of the work at the time was developed using rule-based

algorithms with resolution strategies inspired from either the binding theory or later the centering theory. Two classic anaphora resolution algorithms are particularly important for all of the subsequent work in this area and we detail them here: the algorithm by Hobbs (1978), and the Resolution of Anaphora Procedure (RAP) by Lappin and Leass (1994).

Hobbs' algorithm treats third person referential pronouns only. It traverses the parse trees of the sentences looking for noun phrases of the same gender and number as the anaphor to resolve. The potential antecedents are prioritized according to their categorization, in a way that a subject is preferred to a direct object which is also preferred to an indirect object. It is reported that, in 88.3% of the cases, this algorithm finds the correct antecedent. According to Klappholtz and Lockman (1975), the antecedent is always within the last n preceding sentences, with n often having a value between 1 and 5, $1 \leq n \leq 5$. According to Hobbs' estimation, n is less than one for most of the cases, with 90% of the occurrences in the same sentence. Despite such good figures, Hobbs' algorithm has been criticized because of its assumption of perfect syntactic analysis, since results are computed using parse trees built manually.

The RAP algorithm, on the other hand, treats third person pronouns, reflexives, reciprocals and *pleonastic pronouns, i.e., pronouns without a referent but needed for syntactic reasons*. RAP is based on a series of agreement filters, a binding algorithm which prioritizes arguments according to their categorization –like Hobbs' algorithm– and on salience weighting, a concept of centering theory. It builds on parse trees and identifies referents by analyzing each noun phrase. Each referent has an associated salience value according to the scale in (10), which is updated with every sentence, when the value reaches zero, the potential referent is removed from the list. It should be noticed that the weights listed in (10) are not corpus-based. The authors report 86% of success in resolving anaphora in "edited output of a parser on small, restricted-domain corpus" (Stede 2012, pp. 59-61), and therefore, perfect syntactic analysis. An implementation of the RAP algorithm on the MUC-6 training data by Qiu, Kan, and Chua (2004), however, reached 58% accuracy.

(10) a. Sentence recency: 100
b. Subject emphasis: 80
c. Presentational emphasis (existential sentences like "*there are ...*"): 70
d. Direct object emphasis: 50

e. Indirect object and oblique emphasis: 40
f. Head noun emphasis: 80
g. Non-adverbial emphasis: 50

Moreover, looking forward to access a wider range of text processing frameworks, Kennedy and Boguraev (1996) modify the RAP algorithm so it could work on the output of a POS tagger enhanced with syntactic annotations instead of a parser.

Finally, centering theory principles have also been used by Brennan, Freidman, and Pollard (1987) for AR. In this work, potential centers are identified and ranked according to their type (pronoun or noun phrase) and their grammatical function. The centers are then filtered using constraints to eliminate ambiguity, for instance if an antecedent is proposed for two different pronouns. Finally the centers are ranked according to semantic constraints and the highest ranked one is proposed as antecedent. Brennan, Freidman, and Pollard (1987) and Palomar et al. (2001), referring to German and Spanish respectively, suggest that centering is less straightforward with free word order languages. Since constituents can be displaced to less canonical positions, ranking of entities becomes more difficult.

Other approaches to AR

The decade of 1990 saw a considerable decrease in the research related to AR. From this period, it is to note the knowledge-lean approach of Harabagiu and Maiorano (1999), based on the lexico-semantic knowledge upon which coherence is inferred. Besides, there were the MUC series of conferences (MUC 5-6-7) targeting coreference resolution, including anaphora. Since 2000, however, Spanish researchers contributed to the domain of AR using mainly rule-based methods. They were particularly interested in the resolution of null subjects, also know as zero pronouns. *Null subject pronouns refer to the omission of the subject pronoun which is permitted in some languages*. Ferrández and Peral (2000), for instance, present a null subject coreference resolution system based on restrictions and preferences. The first is intended to create a list of potential antecedents based on c-command, agreement and semantic consistency conditions. The preferences are used to chose the best candidate within the list of potential antecedents. The null subjects are restored as pronoun with the grammatical person features of the verb. Their gender, on the other hand, is only restored when the verb is copulative and

it can be obtained from the predicative object. Later on, the same system is extended by Palomar et al. (2001) to include demonstrative and reflexive pronouns. This enhanced version achieved a success rate of 76.8% vs 75% before. All hand-written rules are dropped and replaced by a preprocessing phase which comprises tagging, parsing and word sense disambiguation in a second extension to the same system, no evaluation is provided though.

Rello and Ilisei (2009) and Rello, Suárez, and Mitkov (2010) further modify Ferrández and Peral's system with a series of heuristics specifying possible contexts where impersonal pronouns occur, in order to differentiate them from genuinely referential null subjects. The issue of distinguishing between regular and pleonastic usages of a pronoun is usually addressed by anaphora and coreference resolution systems. Most of the time, some different types of heuristics are used. The systems by Lappin and Leass (1994) and Kennedy and Boguraev (1996) for instance, rely on pattern recognition of fixed syntactic configurations which are known to involve an expletive 'it', for example *it + to be*. Despite very complete lists of this type of patterns are available, coverage and generality are not assured (Evans 2001).

Coreference resolution

Unlike rule-based systems which rely on heuristics built on deep syntactic knowledge, statistical systems attempt to use as few human-written rules as possible. They rely on several tools for various degrees of preprocessing such as POS-tagging, parsing or named entity recognition. The statistical approach contributes a resolution process at large. Rather than finding antecedents for pronouns specifically (AR), these works are interested in coreference resolution of all referring expressions (mentions) in a text and classifying them into *chains*. The exact preprocessing varies from one system to another but Klenner, Fahrni, and Sennrich (2010) and Klenner, Tuggener, et al. (2010) strongly emphasize the importance of the quality of the tools used during the pre-processing, as it is crucial for correct coreference identification in a fully automatic manner. However, as they explain, each individual tool introduce errors, decreasing the overall performance.

The resolution strategy proposed by Soon, H. T. Ng, and Lim (2001) is one of the most influential works. Their system first identifies all mentions –potential antecedents and anaphors together– and decides whether or not two mentions are coreferential according to a pairwise classification paradigm based on the features listed in (11) below. This

system achieves a recall of 58.6% and a precision of 67.3% on the MUC-6 corpus (Grishman and Sundheim 1995); 56.1% recall and 65.5% precision on the MUC-7 corpus (Grishman and Sundheim 1995).

(11) For any pair (i, j) of markables:

a. distance between the markables
b. whether a markable i is a reflexive, personal or possessive pronoun
c. whether a markable j, is any pronoun
d. whether the agreement features of i and j match
e. whether j is a definite NP
f. whether j is a demonstrative NP
g. whether i and j agree in number
h. semantic class agreement (semantic classes include: female, male, person, organization, location, date, time, money, percent and object)
i. whether i and j agree in gender
j. if i and j are both proper names
k. whether i is an alias of j
l. whether j is an apposition to i

Surveys such as those conducted by Strube (2007), Stoyanov et al. (2009), V. Ng (2010), Poesio, Ponzetto, and Versley (2010), and Mitkov (2010) provide detailed information on the systems developed recently and review some of the problems of the state-of-the-art in the coreference and anaphora resolution field. Table 2.1 shows a comparison of the systems mentioned here. We are aware that the systems listed differ in their structure, the approach used and the corpus employed in the evaluation. However, such a comparison remains useful for a general overview of the performance reached by the resolvers. Besides, it is indicative of the intuitions followed by researchers on the type of referring expressions to resolve. Finally, there are AR systems freely available. Some widely cited ones in the literature include an implementation of RAP by Qiu, Kan, and Chua (2004), MARS (Mitkov, Evans, and Orăsan 2002), BART (Versley et al. 2008), RECONCILE (Stoyanov et al. 2010), Stanford CoreNLP (H. Lee et al. 2011), and, more recently, CORT (Martschat and Strube 2015).

System	Type	Tested on	Referring Expressions	#	Accuracy	Language
Hobbs (1978)	rule-based, AR	Parse trees manually built	he, she, it, they	300	88.3%	en
RAP by Lappin and Leass (1994)	rule-based, AR	Parse trees manually built	he, she, it, reflexives, reciprocals	560	86%	en
RAP by Qiu, Kan, and Chua (2004)	rule-based, AR	MUC-6 corpus (Grishman and Sundheim 1995)	he, she, it, reflexives, reciprocals	n/a	58%	en
Kennedy and Boguraev (1996)	rule-based, AR	Random selection of 27 texts	he, she, it, reflexives, reciprocals	306	75%	en
Ferrández and Peral (2000)	rule-based, AR	Blue Book and Lexesp	he, she, it, null subject pronouns	228	75%	es
Soon, H. T. Ng, and Lim (2001)	statistical, CR	MUC-6 corpus (Grishman and Sundheim 1995) MUC-7 corpus (Grishman and Sundheim 1995)	noun-phrases noun-phrases	n/a n/a	62.6% (F1) 60.4% (F1)	en
Mitkov, Evans, and Orăsan (2002)	statistical, AR	Computer manuals	he, she, it, they	2,263	59.39%	en
Klenner, Fahrni, and Sennrich (2010)	statistical, CR	Computer manuals	noun phrases	n/a	58.01 (F1)	de

Table 2.1: Comparison of selected anaphora resolution (AR) and coreference resolution (CR) systems. Unless explicitly specified, we report accuracy measures.

2.2.3 Machine translation of pronominal anaphora

As mentioned earlier, discourse knowledge has been claimed to be valuable for several NLP applications, MT among them. To give an example, discourse parsing would allow to disambiguate all pronouns based on the right frontier constraint as pointed out before, although a system with the required performance to achieve this does not exist yet. A growing body of researchers is keen on studying discourse and MT. A marked interest in improving the MT of pronouns emerged around 2010, partially as an effect of the progress of MT quality using statistical models (SMT).

The shortcomings in the quality of current MT output are partially due to the limitations of the essential assumption of sentence independence. Researchers working on discourse and MT have shown that some aspects of MT quality can improve if discourse knowledge, i.e., processing of phenomena beyond the sentence level, is taken into account (Hardmeier 2014; Meyer 2014; Guillou 2016). In this context, the first works interested in improving the MT of pronouns explored the possibility of integrating AR into MT. These propositions are based on annotation projection strategies, which showed very modest results. Such strategies proved even less fruitful when the languages involved are typologically distant such as Czech and English and the translation quality is only moderate (Guillou 2012).

Integrating AR in machine translation

One of the first proposals to aid pronoun translation found in the literature is to incorporate an AR system at some point during the translation process. The difficulty with this strategy is that an AR system operates at a monolingual level, while MT necessarily needs to account for the cross-lingual properties of the source and target languages involved. The agreement features of a pair antecedent-pronoun may change when they are translated into another language, as illustrated in (1). An AR system would provide the knowledge about the link between *bike* and *it* in (1a), but does not provide any clue on the translation of the pronoun in French. If one chooses a translation like (2a), the pronoun must be masculine, whereas a translation like (2b) implies a feminine translation.

(12) a. ENGLISH He left *the bike* outside the house without locking *it*.
b. FRENCH Il a laissé *le vélo* hors de la maison sans *le* verrouiller.
c. FRENCH Il a laissé *la bicyclette* hors de la maison sans *la* verrouiller.

This solution is therefore implemented in two steps. In the first step, the resolution system is used in the source text so antecedents are identified. The text is then translated and this first translation is used for extracting the gender of the antecedent in the target text. Next, the source text is annotated with the gender information. Finally, the text with the annotations is translated a second time. Both, Le Nagard and Koehn (2010) and Guillou (2012) use this technique. Le Nagard and Koehn (2010) use both, the Hobbs (1978)'s and Lappin and Leass (1994)'s coreference resolution algorithms on a English - French SMT system. After the second translation, results of correctly translated pronouns (69%) did not vary much from the baseline (68%). They attribute their results to the quality of the coreference resolution algorithms. Using the same approach, Guillou (2012) controls for the coreference resolution quality using gold annotations instead. She works on the English to Czech pair but obtains little improvement as well.

Working with an English to German system, Hardmeier and Federico (2010) address the two step problem feeding the decoder one sentence at the time and putting it in a queue after being translated. If a pronoun is detected, then the translation of the antecedent is recovered from the queue using the alignment information, and the sentence containing the pronoun passes into the decoder along with the information about the gender of the pronoun's antecedent. They obtained little overall improvement. However, after a manual evaluation, they found that the translation of pronoun *it* was much better by their system than the baseline.

These last four pieces of research exploit the idea of using parallel data information (non-available to standard AR or CR systems) to support the translation, accomplishing the resolution as a byproduct of the process. The idea of taking advantage of some information from one language to complement what is missing in the other was introduced by Mitkov and Barbu (2002). Their intuition is that difficult cases for standard AR systems in English, could be easily disambiguated using parallel corpora aligned at the word level. They give the example repeated here as (13). In (13a), the pronoun *it* is a difficult case for AR systems because *John* and *cassette* are both initially considered potential antecedents which are syntactically and semantically more prominent than *videoplayer*, the first is the subject and the second is the direct object. *John* can be easily discarded as antecedent on semantic grounds, since it is a human subject, yet, *videoplayer* is realized as propositional phrase, a syntactic type heavily penalized by most AR systems. Because French is marked for gender, the authors argue that cases similar to the one presented in example (13) could be disambiguated by looking at the French word-aligned

Pronoun translation correspondence	**#**
Pronoun to pronoun	241
English pronoun to French noun phrase	24
English noun phrase to French pronoun	68
Nothing to French pronoun	91
English pronoun omitted from French translation	16

Table 2.2: English-French translation mapping of 390 pronouns reported by Mitkov and Barbu (2002).

translation. In this case, because in French *cassette* is feminine, and both *magnétoscope* and the translation of the pronoun itself *le* are masculine they can be matched safely as coreferring.

(13) a. ENGLISH John removes the cassette from the videoplayer and disconnects it.

b. FRENCH Jean éjecte la cassette du magnétoscope et le débranche.

From a corpus study using the parallel corpus, the same authors observed that pronouns are not necessarily translated as pronouns on a one-to-one basis (Mitkov and Barbu 2002, pp. 202-203):

> Some of the pronouns occurring in English were completely omitted in French, replaced by full noun phrases or replaced by other types of anaphors whose resolution was not tackled in the project (for example, demonstratives). Similarly, some English noun phrases were replaced by pronouns in the French translation, whereas a few additional French pronouns were introduced even though they did not have a corresponding pronoun in the English text.

In their corpus, there are 281 English pronouns while for the same French text there are 390. They report the mapping summarized in Table 2.2. A similar finding is reported by Weiner (2014) for the translation from English to German. In this work, the author does a preliminary evaluation and finds that around 70% of the source pronouns that are aligned to target pronouns are translated correctly by a baseline system. However, looking at types of pronouns individually, he reports that only 47.6% of *it* pronouns are translated correctly in a newswire corpus, caused by a bias of the SMT towards the neuter translation.

These figures contradict the intuition that a pronoun in the source language must correspond to a pronoun as well in the target language. In other words, the problem of pronoun translation is not just a matter of choosing a correct corresponding pronoun of the same category. We will explore this issue again in Chapter 4. Furthermore, the distribution of pronouns and their difficulties of translation are specific to the source and target languages of interest and the text genre as well (Russo et al. 2011; Scherrer et al. 2011).

Cross-lingual pronoun prediction

Combining a coreference or anaphora resolution system with a SMT system resulted in a unsatisfactory solution to the problem of pronoun translation. On the one hand, the translation needs to be accomplished in two passes; on the other, AR and CR systems rely on heavy preprocessing, with several sub-tasks which are themselves imperfect, introducing errors. Besides, most existing AR and CR systems are adapted to work on English.

As an alternative to an AR system, Popescu-Belis et al. (2012) suggest to predict the target translation of the source pronouns using source side context information only. Working with a manually annotated corpus of 400 instances of *it* and their translation into French, they present a pilot experiment which achieves only moderate success.

Cross-lingual pronoun prediction is a classification approach to estimate a pronoun's translation directly, without generating a full translation of the segment containing the pronoun. The task is defined as a fill-in-the-gap task: given an input text and a translation with placeholders, replace the placeholders with pronouns. It is an attractive method to approach the pronoun translation problem because it limits it to a small number of classes. Besides, unlike full MT, it is a setting where both source and target language data are available (excepting the target pronoun) during training and testing time. Both languages can be analyzed to create features which encode different types of information as in a standard classification approach, potentially providing the means to understand the different aspects involved in pronoun translation.

The prediction approach was noticeably developed and formalized by Hardmeier (2014) in the form of a cross-lingual pronoun prediction task. This task was first introduced as a shared task at the DiscoMT 2015 Workshop (Hardmeier, Nakov, et al. 2015) and

Subtask	Source pronouns	Target pronouns
EN-FR 2015	it, they	ce, elle, elles, il, ils, cela, ça, on, OTHER
EN-FR 2016	it, they	ce, elle, elles, il, ils, cela/ça, on, OTHER
EN-DE 2016	it, they	er, sie, es, man, OTHER
FR-EN 2016	elle, elles, il, ils	he, she, it, they, this, these, there, OTHER
DE-EN 2016	er, sie, es	he, she, it, you, they, this, these, there, OTHER

Table 2.3: Source and target pronouns defined for the 2015 and 2016 shared tasks on cross-lingual pronoun prediction. The OTHER class is a catch-all category for translations as lexical noun phrases, paraphrases or nothing at all (when the pronoun is not translated).

was repeated in 2016 with some modifications[4] (Guillou, Hardmeier, Nakov, et al. 2016). The 2015 task was focused on English into French translation only, while the 2016 series include English-German and considers both language pairs in both directions. Another difference is that the 2016 task provides lemmatised target-side data, approximating a more realistic MT-like scenario. A contrastive summary of the classes for each language pair in both shared tasks is presented in Table 2.3. These two shared tasks gave visibility to the problem and consequently many papers on the subject were published in the last two years. Many classification approaches have been used for cross-lingual pronoun prediction, including n-gram models, linear classifiers, neural classifiers, among others.

Linear and kernel methods

Using a maximum entropy classifier, Wetzel, Lopez, and Webber (2015) include as features the closest NP antecedent candidate in the target language as identified by the Stanford CoreNLP (H. Lee et al. 2011) in the source language and local context (three tokens around the source and target pronouns). Moreover, they include the probability of *it* being a pleonastic pronoun produced by the system NADA (Bergsma and Yarowsky 2011). They also include the prediction of a 5-gram language model queried with the concatenation of three words before the pronoun in their features. Wetzel (2016) slightly modifies the features in this work for English into French to be compatible with English into German pronoun prediction. Besides the token context features, he includes the target context POS-tags and their n-gram combination. The NADA probability is included, as well as the LM prediction but this time queried with the complete sentence. The author note that including the LM features did not help the classifier. Additionally,

[4]The author is currently coordinating the 2017 edition of the shared task.

he compares the results obtained with those obtained with a conditional random fields (CRF) classifier in which the pronouns in the sentence are modeled as sequences. The results of the maximum entropy classifier systematically outperformed those of the CRF.

Novák (2016) proposes a multiclass variant of the logistic loss and stochastic gradient descent optimization as implemented in the Vowpal Wabbit toolkit for English-German in both translation directions. For feature extraction, he parses the source text using Treex framework (which also produces semantic roles), a POS-tagger and a dependency tagger as well. Coreference links to potential antecedents are obtained by a combination of the Treex parser and the BART toolkit. From this, he extracts the gender in the target of the projected noun antecedent in the source. He also includes the probability of *it* being pleonastic as computed with the NADA system.

The system by Tiedemann (2015) is built using a support vector machine (SVM) classifier (Fan et al. 2008) and relies heavily on local context features in a bag-of-words manner. It ranked first among the submitted systems to the 2015 task, outperformed only by the baseline. He found that right (source) context is more important than left source context, but overall target context is more informative, with an an optimal (target) window of two words to the left and three to the right of the pronoun of interest. Additionally, he uses the gender and number of the preceding noun phrases (these are taken from the aligned tokens to English determiners *a*, *an*, *the*, *those*, *this*, *these* and *that*). This system is compared with a position sensitive approach as well. The first produce higher F-scores for small windows but has lower accuracy than position-sensitive models. According to the author the system works well with the OTHER and *ce* classes, as well as the masculine plural *ils*. Most problems can be found in the predictions of the female pronouns *elle* and *elles* and the demonstratives *cela* and *ça*, all of which have low frequency. The same system performed significantly less well in the setting with lemmatized target data (Tiedemann 2016).

Stymne (2016) also built a SVM classifier and ranked second in the 2016 task. She experiments with a wide range of features, including the source pronouns, the source and target-lemma context of up to four preceding nouns and their POS-tags, and gender and number for the target side proper names. She also includes target POS-tags n-grams of several sizes, the dependency heads of the source pronous, target LM scores, the number of tokens that a pronoun is aligned to, the length ratio between the source and target sentence where the pronouns occurred and the relative pronoun positions (beginning of sentence or not). Interestingly, the best features turned out to be the local context and

the dependency links, while the features related to the LM and the preceding nouns hurt performance as also reported by Wetzel (2016).

Classification approaches focused on the target language

Bawden (2016), on the other hand, builds a random forest classifier for the English-French pair. Contrary to most systems, she extracts most of her features exclusively from the target language. She runs a modified parser on the target text (since the target is lemmatized) enriched with gender information taken from the Lefff dictionary (Sagot 2010) (including therefore ambiguity for many forms). She includes the baseline probabilities of the most probable class and the next two most probable classes as well. In addition, she uses the gender and number of antecedent as identified by the Stanford CoreNLP system, expletive detection heuristics, and, syntax-based context features (class of words in the proximity of the pronoun). Interestingly, she does a manual evaluation which confirms the low quality of coreference resolution system for French: "Of 237 pronouns of the form *il*, *elle*, *ils* or *elles*, 194 were anaphoric with a textual referent. The correct referent was provided in only 52.6% of cases, the majority being for the masculine plural class *ils*. The tool also often fails to predict impersonal pronouns, erroneously supplying coreference chains for 18 impersonal pronouns out of 25" (Bawden 2016, p. 566).

Another system focusing on the target language is the one presented by Luong and Popescu-Belis (2016b). They propose a combination of a target side pronoun LM and heuristics. The pronoun LM is trained on sequences of gender of the nouns in the target. Their idea is to capture the likelihood of a pronoun given the gender and number of the nouns or pronouns preceding it. For the actual prediction, they first apply several heuristics to discriminate whether the predicted pronoun belongs to the *on* or OTHER classes, which indicates that the uncertainty of the translation collected in the OTHER class is not modeled by the LM. The pronoun LM is then used to score all remaining possible candidates and the one with the highest score is selected. This approach proved much less effective than all the previous work, although later versions offer improved results Luong and Popescu-Belis (2016a).

Neural classification

As it has happened in the MT domain in general, deep learning models have gained followers for pronoun prediction and translation. For this particular task, the work of Hardmeier, Tiedemann, and Nivre (2013) is one of the first to use neural models. They build a feed-forward neural network classifier trained on features from both the source and the target language data. From the source, they extract the pronoun context (3 words to the left and 3 words to the right), while from the target, they extract the words aligned to the head identified as potential antecedent in the source language using the coreference resolution toolkit BART (Versley et al. 2008). This 'target antecedent' is retained if it matches in gender to the pronoun to predict. Additional training features are extracted using the coreference resolution system but the explicit links to the antecedent are not used. This classifier is compared to a baseline maximum entropy classifier trained on the same set of features. The neural networks classifier gained 0.027 points of precision and 0.054 points of recall over the baseline and it is argued to perform at the level of state-of-the-art coreference resolution systems, performing particularly well with low-frequency classes such as the feminine pronoun *elles*.

In a later stage of the same work, the classifier is combined with an SMT system using the Docent decoder (Hardmeier, Nivre, and Tiedemann 2012). Docent is a document-level decoder which functions very much like standard phrase-based SMT, but scoring documents as a whole. In addition, it may include scoring functions that are sentence-internal or that go beyond sentence boundaries. For evaluating this model, both the language model and the SMT system are trained with placeholders instead of the pronouns aligned to either *it* or 'they'. This method isolates the effect of the classifier since it blocks any effect of the baseline translation. In other words, pronouns are always predicted, avoiding to confuse correctly predicted pronouns with pronouns that otherwise would have been translated correctly by a normal training. Results for pronouns in a Newswire corpus degraded a bit and those for the TED corpus (Cettolo, Girardi, and Federico 2012) were minimally improved (0.002 F1 difference between a baseline and the predicted systems) (Hardmeier 2014).

In Hardmeier (2016), the author re-adjusts to work in the lemmatized setting of the 2016 shared task. The source pronoun context features remain the same, while the target context features are lemmatised. The target words aligned with the heads of the potential antecedents are used as well, but the coreference resolution system used to find them

is the CORT toolkit (Martschat and Strube 2015). The coreference links are included as well. The output of the classifier is combined with a LM fed with source pronoun information. Contrary to the original work, the final system is not good with rare pronoun classes, in particular *elles*, as pointed out by the author.

Pham and van der Plas (2015), on the other hand, use a neural network classifier and word distributed representations of the English and French context (three words before and three words after each English and French pronoun) as features. In addition, French morpho-syntactic information produced by parsing the target with Morfette (Chrupała, Dinu, and van Genabith 2008) is added as features. As other systems mentioned so far, they use the Stanford CoreNLP toolkit (H. Lee et al. 2011) to obtain antecedent candidates in the source language (the closest noun phrase in the coreferential chain containing the pronoun) and use the alignment information to obtain the target language token. However, they note that coreference information did not added much knowledge to the model.

A feed-forward network model trained on distributed representations of the source and target context is also presented in Callin, Hardmeier, and Tiedemann (2015). They use a context window of four words instead of three and their POS-tags. Consistently with some of the works presented until this point, they report that the right context is more decisive that than left context for the classifier, which makes sense since the predictions correspond to subject pronouns, but contradict the intuition that the crucial information come from the antecedents which are most likely to the left.

Contrary to the neural systems described before, Dabre et al. (2016) and Luotolahti, Kanerva, and Ginter (2016) use recurrent models, which have been proved effective for predicting sequences. Dabre et al. (2016) build a recurrent neural network with stacked GRU units and attention mechanism. For this system, all the words in the sentence either to the left or to the right of the pronoun in both the source and target languages are used as context features. Luotolahti, Kanerva, and Ginter (2016) built the best ranked system for the 2016 shared task. The system is based on two stack levels of GRU units and it relies almost uniquely on context. Other than representations of the source pronouns, its input contains a fixed window of 50 tokens, reading away from the pronoun to be predicted, to the left and the right, both for the source and the target language. It includes a weighted loss which penalizes classification errors on low frequency classes.

Discriminative word lexicon

Using rather different strategies, Weiner (2014) proposes two approaches to include pronoun predictions of a classifier into a SMT system for the English to German translation. His first method targets pronoun translation as word disambiguation problem based on Discriminative Word Lexicon (DWL), a model for the occurrence of individual words in the translation output. DWL intervenes when computing the phrase table of the SMT system, but instead of being computed on phrases, it is computed on single words. It models the probability of the set of target words in a sentence e given the set of source words f. It consists of individual classifiers for each word in the target e_j in the translation of a given source sentence f (Mauser, S. Hasan, and Ney 2009). This technique produces no improvement in the translation but using the same method, Herrmann, Niehues, and Waibel (2015) reported an increase in pronoun prediction accuracy of between 5% and 9%. Weiner (2014)'s second method is based on DWL as well. The model includes not only the set of source words as features, but also the set of source n-grams and the target antecedent. For extracting the target antecedent he used the implementation of the RAP algorithm by Qiu, Kan, and Chua (2004) on the English side of the corpus and word-level alignments for its German correspondence. The results of this second approach were slightly positive but not significant.

Full MT with focus on pronoun translation

Other than cross-lingual pronoun prediction, there are some works comprising a full translation pipeline, but focused on the evaluation of pronouns only. A full translation subtask was also included in the 2015 DiscoMT shared task. It required participants to submit the complete English text translated into French, although only pronouns were evaluated. Two of the submitted systems used the Docent (Hardmeier 2014) document level decoder with different strategies for handling pronouns. Tiedemann (2015) used an n-gram language model over the POS-tags of words linked to English pronouns and determiners. Hardmeier (2015), on its part, uses a neural network classifier for pronoun prediction fed with context and antecedent information from the Stanford CoreNLP toolkit (H. Lee et al. 2011). Both systems performed just under the baseline. Note that, as in the pronoun prediction track, none of the submitted systems beat the phrase-based system baseline.

Post-processing

Improving pronoun translation can also be seen as a post-processing corrective task. Guillou (2015), for instance, proposed an automatic post-editing system trained on the ParCor corpus (Guillou, Hardmeier, Smith, et al. 2014). The system includes pleonastic and referential pronoun identification using NADA (Bergsma and Yarowsky 2011). Pleonastic pronouns are left untreated while antecedents for referential pronouns are found using the Stanford CoreNLP toolkit (H. Lee et al. 2011). The gender and number of the French word aligned translation to the antecedent are taken as values to propose a corrected pronoun. If an antecedent is not found, a default value *il* or *ils* is proposed. Otherwise, Luong, Miculicich Werlen, and Popescu-Belis (2015) propose to correct the subset of French pronouns *il*, *ils*, *elle* and *elles*, based on their grammatical function, subject or object, since each corresponds to different possibilities in French. Their correction is based on a score combining the decoder's score from a SMT system search graph and the general accuracy score of a coreference resolution system. The selected pronoun is the one that maximizes the combined scores of these two criteria.

Weiner (2014) also presents two post-processing strategies for a German-English system. In the first, he creates a list of English pronoun-antecedent pairs using the RAP implementation by Qiu, Kan, and Chua (2004), then he extracts the German equivalents from the bitext. Using POS-tagging, the agreement between the German pronoun-antecedent is checked and corrected if necessary. This method improves the pronoun translation accuracy in 5.6% but BLEU scores[5] do not change significantly, as expected. In the second strategy, he uses a pronoun-antecedent pairs list again. Separately, a list with the 300 n-best hypotheses per sentence is built. Afterwards, the pronouns in each hypothesis are checked against the first list and the hypothesis with the highest translation probability is preferred. This second method did not produce significant changes, since correct anaphora-antecedent pairs are almost never generated among the 300 n-best hypotheses.

Tectogrammatical framework

Finally, recent developments in the Prague Dependency Treebank (PDT) (Kučová and Hajičová 2005) concerning coreference annotation (Nedoluzhko, Mírovský, and Novák

[5]The BLEU score is an automatic measure of precision computed on the comparison between the translation produced by the system and a human reference translation (Papineni et al. 2002). Details are given in Section4.4.

2013) have led to research in MT of pronouns for the English - Czech translation. The PDT's annotation follows the principles of the Prage Tectogrammatics Theory, which represents the semantic structure of the sentence (Lopatková, Plátek, and Sgall 2008). In this framework, Novák (2011) and Novák, Nedoluzhko, and Žabokrtský (2013) describe a transfer MT system with a rule-based pronoun resolution strategy based on default choices. Since the PDT also has a parallel English text, each *it* is word-aligned to its translation, creating a corpus which has also been used as gold training data for pronoun prediction classifiers. The features for these classifiers are extracted exclusively from the source text (English) and are inspired by grammar rules pertinent in disambiguating the different translations of 'it'. Despite experimenting with different machine learning algorithms (binary logistic regression, maximum entropy, k-nearest neightbors, decision trees, SVM), results were consistently biased towards the majority class and the maximum performance obtained of around 70% accuracy is similar to the work on English-French pronoun prediction.

Assessment

Aiming at text-level translation, all the pieces of research for pronoun prediction presented can be viewed as efforts in a more accessible task than full translation. The features fed to many of the systems include inter-sentential dependencies such as antecedents and anaphors, and also larger text spans than the current sentence. Explicit antecedent information found using a coreference resolution system is part of the features of some systems, although it does not seem to ensure better performance. There does not seem to be a clear trend in this respect. However, the surrounding source and target context is exploited by all the systems and this feature is clearly important to the point of even providing some information about the antecedent. It should be noted that the best performing systems to both tracks of the 2015 shared task, cross-lingual pronoun prediction and full translation, were the baselines: a 3-gram LM and a standard phrase-based system with the same LM. Nevertheless, context alone is not enough information to predict low frequency classes less driven by context and more by semantic and morphological information such as *elle* and *elles*. More (linguistic) understanding of the role of language models for pronoun prediction and translation is needed. Also in this line, as pointed out by some of the authors, neural models capture a lot more of the relevant information and make better predictions, but it is not entirely clear how.

In this work, we investigate the informativity of context features as opposed to morphological and syntactical features for pronoun prediction (Chapter 5). Moreover, we are interested in the role of source pronoun function for target pronoun prediction. With pronoun function, we refer to Guillou (2016)'s distinction between pronouns with a nominal anaphoric, eventual anaphoric and pleonastic function (Chapter 6). In what follows, we review the existing relevant work concerning temporal reference and the machine translation of verbal tenses.

2.3 Temporal reference

Temporal reference enables the indication of time using grammatical means. It places a state or action, i.e., an event,[6] in a particular point in time or moment of speech (De Beaugrande and Dressler 1981; Zufferey and Moeschler 2012). Several linguistic theories exist that try to account for the grammatical means and semantic cues which permit the temporal interpretation of the different tenses. Reichenbach, for instance, argued for the identification of three different times for an appropriate account of tenses: the utterance time, the reference time and the event time. Kamp (1979), on its part, introduces the idea that the type of event has an influence on the general expression of tense and therefore on its interpretation (Garnham 2001).

Temporal reference is not strictly parallel to pronominal reference in the sense that tense morphemes do not point to a specific antecedent, as pronouns do; however, tense is considered an anaphoric category (Reichenbach 1947; Partee 1973; Partee 1984; Moens and Steedman 1988). The referent for a verbal tense can be a temporal adverbial expressions such as *at five o'clock last Saturday*, *this morning* or the moment of speech, as in the case of the present tense.

The notions of tense, aspect and mood are the means through which the different temporal interpretations of an event surface or grammaticalize in a language. There is overlap between the extent of these concepts and their definitions are somewhat inconsistent between different linguistics traditions. Indeed, tense and aspect are particularly difficult categories to describe, specially in the context of cross-linguistic comparison, as

[6]The term *eventuality* as well as *situation* are generic terms which include all types of verbs. The term *event* is often used as a synonym, especially in computational approaches. In the linguistic literature, however, the term *event* does not include verbs considered as *states*.

languages differ in their way of expressing the temporal location of events (Dahl and Velupillai 2013). Germanic and Romance languages have a large grammaticalization of the notion of tense while much less of the notion of aspect. In the case of Slavic languages, the notion of aspect is more prominent and therefore grammaticalized. Languages such as Chinese, on the other hand, do not have specialized temporal verbal morphology, but use other means such as particles and adverbials.

2.3.1 Automatic classification of verbal tenses

Within the language technology domain, tense discrepancies between the languages have been investigated through tense prediction tasks. The studies described by Ye, Fossum, and Abney (2006) and Ye, Schneider, and Abney (2007) are interested in the temporal mismatches in automatically translated texts from English to Chinese. The problem in this pair of languages is that, on the one hand, English is marked for tense but less for aspect (with the exception of the progressive marking); on the other hand, Chinese aspect –if marked– takes the form of a separate word which aligns poorly with English tensed verbs, and so the aspectual information is dropped from the translations. As a result, instead of producing (14) SMT systems produce the sentence (15), using the infinitive form of the verb and, in this case, with a different lexical choice.

(14) Wo ji le yi feng xin gei ta.
1st send PERF one QUA letter PP 3rd
'I *sent* him a letter.'[7]

(15) Wo xie yi feng xin gei ta.
1st write one QUA letter PP 3rd
'I *write* him a letter.'

The first study (Ye, Fossum, and Abney 2006) examines the utility of features of different nature for predicting the correct tense of English verbs inadequately translated from Chinese. In particular, latent features, inspired by how humans interpret temporal relations in text, are tested and compared with surface features such as the use of quoted speech, the clause syntactic type, presence of temporal adverbs, distance between the previous and current verbs. Examples of latent features include telicity, punctuality and

[7] 1st - first personal pronoun, 3rd - third personal pronouns, PERF - particle of completed and perfective eventuality, PP - preposition (*to/for/for the benefit of*), QUA - quantifier

temporal relations of the verbs.[8] A classifier of present, past and future tense, trained on 2,500 verbs and on surface features and a classifier trained on latent features both underperformed in comparison with one trained on both types of features, reaching 83.4% accuracy. The authors argue, consequently, in favor of the value of using latent features for tense prediction and for NLP in general.

In the second study, their objective is to predict the appropriate Chinese aspect marker and to insert it in the Chinese translation. A classifier is trained on 2,723 verbs annotated with one of four possible Chinese aspect markers. They obtained a general accuracy of 77.25%. Contrary to the previous study, however, a feature utility ranking showed a low impact of the aspectual features of punctuality and telicity.

2.3.2 Machine translation of verbal tenses

In these previous studies the classification results were not embedded in a SMT system and the classifier classes were the actual verbal tenses. In contrast, Meyer, Grisot, and Popescu-Belis (2013) use classification as a means of enhancing a SMT system with knowledge about *narrativity* in order to produce better tense choices at translation time. Narrativity is a pragmatic property triggered by tense and refers to determining the status of the temporal relations holding among events. Two cases are possible: *narrative* and *non-narrative* usages of a verbal tense. A narrative usage points to the case when the two events are temporally linked (with both forward and backward temporal inferences). Non-narratives usages point to the case when events are either not temporally linked or they occur simultaneously.

In their paper, Meyer, Grisot, and Popescu-Belis (2013) built a classifier, which was trained on a small manually-annotated corpus with narrativity, to generate narrative and non-narrative disambiguation labels for the English simple past (SP) verbs of a large parallel corpus. In other words, they classify the SP verbs of the SMT training data into narrative or non-narrative instances. With this second corpus, they built a SMT system using a factored model of translation. This system gained 0.2 BLEU points when compared to a baseline system lacking the disambiguation labels. The authors note two

[8]Telicity, or *aktionsart* (Vendler 1957), is a classification based on potential endpoints. The event of 'running a marathon', for instance, has an inherent endpoint (Declerck 2007). Punctuality specifies if an event is associated with a specifif point in time as opposed to events with a duration in time. Typical punctual verbs include breaking, blasting and jumping (Engelberg 1999). Temporal relations between two events include precedence, succession, inclusion, subsumption, overlapping or no temporal relation at all.

shortcomings in their method. Firstly, the classification results are rather moderate (F1 = 0.71), since narrativity is hard to infer from surface clues. Secondly, they note a problem with the identification of the SP verbs in the large corpus, in particular when used in the passive voice (for instance, instead of *was taken*, they only detect *was*).

Following the example of Meyer, Grisot, and Popescu-Belis (2013), in our work we built a classifier trained on a small manually-annotated corpus and then used the classifier to annotate a large corpus for training a SMT system. In our study, we use the property of *boundedness* (Chapter 8). Each SP instance is annotated with a *bounded* or *unbounded* label and these labels are then used as disambiguation markers. Compared to *narrativity*, *boundedness* is more likely to be correctly learned by a classifier on the basis of surface clues and linguistically-informed features. Finally, we use a more sensitive method to identify English SP verbs either in the active or passive voice.

Meyer, Grisot, and Popescu-Belis (2013), Loáiciga, Meyer, and Popescu-Belis (2014) and Loáiciga and Grisot (2016) present the only existing work on statistical machine translation of verbal tenses between English and French. Most of the work on machine translation of verbs concerns the translation between Chinese and English. In Chinese, the grammatical aspect markers for *perfective* and *imperfective* are optional. Therefore, Chinese verbs are underspecified when compared to English, and what in English would correspond to present and past tenses, for example, are hard to distinguish in Chinese, compromising the quality of translation. Addressing this problematic, Olsen et al. (2001) report probably the work most closely related to our own. The particular architecture of their system (interlingua model) allows them to obtain reliable semantic information associated with each verb. This information includes primitives (GO, BE, STAY, ...), types (Event, State, Path, ...) and fields (locational, temporal, possessional, identificational, perceptual, ...). Using this information and some heuristics which exploit additional clues from the sentence such as adverbs, they implement an algorithm that identifies telic Chinese verbs. Their hypothesis is that Chinese sentences with a telic aspect will translate into English past tense and those without the telic aspect as present tense.

Their system is tested on a 72 verb test set matched against a human reference translation. Results are given in terms of accuracy or correct translations. While the baseline system obtained 57% correct translations, a second system which uses the telic information of verbs obtains 76% correct translations. Furthermore, a third system built using the telic information along with other linguistic information such as grammatical aspect

and adverbials obtained 92% accuracy. Contrary to our framework, this system is highly deterministic, with a fixed correspondence *+telic* → *past*, *–telic*→ *present* which might be incorrect in other language contexts. Besides, the identification process of telic verbs relies heavily on their particular system's lexicon, making it difficult to implement in a SMT setting.

In the same context of Chinese to English translation, Gong et al. (2012b) propose a method to reduce tense inconsistency errors in MT. Based on sequences of the tenses in a sentence (e.g., present, present, past), they build an n-gram tense model of the target English side of the corpus. At decoding time, when a hypothesis has covered all source words, the tense of the main verb in the current sentence is predicted first and then the complete tense sequence of the previous sentence. With this information the translation hypothesis is re-scored a, including the weight of each predicted tense found by minimum-error-rate training (MERT) (Och 2003). They gain 0.57 BLEU points using the tense sequence information, 0.31 using the main tense information, and, 0.62 using the combination of both.

The same authors report on a follow-up study (Gong et al. 2012a) which additionally uses information concerning the source language Chinese to extract the features given to the classifier. This classifier is trained to assign one of four tense labels to Chinese verbs before translation. Each of these labels has an a associated probability, and the highest one is retained. As before, during decoding time, this probability is fed to the SMT system and the hypothesis translations are re-ranked. They obtain a BLEU score improvement of 0.74 points.

Finally, as part of a study mostly interested in reordering English verbs when translating into German, Gojun and Fraser (2012) report a pilot experiment concerning verb tense disambiguation. They trained a phrase-based SMT system using POS-tags as disambiguation labels concatenated to English verbs which corresponded to different forms of the same German verb. For example the English *said* can be translated in German using a past participle *gesagt* or a simple past *sagte*. This system gained up to 0.09 BLEU points over a system lacking the POS-tags.

Assessment

The subject of temporal reference enjoys less empirical work concerning MT than the subject of pronominal anaphora. Similarly to pronouns that refer back to an antecedent

to find their meaning, verbal tenses refer back to several elements in the sentence to find their interpretation. Existing work on MT and verbal tenses has focused on the study of tense and aspect as disambiguation criteria for the translation between English and Chinese and English and French.

In this thesis, a large-scale corpus study on the translation of the verbal tenses between English and French is presented. This corpus study results in a parallel corpus annotated automatically with tense information. The annotation method is described as well. Afterwards, following existing work on the translation of verbal tenses, we focus on the translation of the English simple past into French, presenting experiments which exploit existing ressources and context information. These two points are developed in Chapter 7 and Chapter 8 respectively.

2.4 Conclusion

A discourse forms a comprehensible text when the patterns formed by its segments (clauses, sentences or paragraphs) convey meaning and this meaning is more than what each segment conveys alone. For the transmission of meaning, a text exploits several language features. Pronominal and temporal reference are cohesive devices that contribute to a text coherent interpretation. The study of cohesive devices has been focused on the mechanisms for adverbial, temporal and nominal reference, addressing the grammatical categories of adverbs, verbs, and nouns and pronouns respectively. In this work, we focus on verbs and pronouns and their problems and treatment in MT.

Concerning pronominal reference, there is a growing body of researchers interested in the subject. Humans rely on a complex interaction of linguistic and world knowledge which makes the task of interpreting pronouns relatively trivial, even in highly ambiguous contexts. The formalization of this process has not been without its difficulties. SMT researchers have taken an interest in the problem with a cross-lingual perspective. Benefiting from the many AR and CR systems available, they have tried to combined these systems with a MT system, obtaining a few positive results. Current methods address the problem of pronoun translation as a cross-lingual pronoun prediction task. Temporal reference, on the contrary, has been much less studied and has been modeled almost exclusively as a tense prediction task.

Chapter 3

Tools

In this chapter, we outline the tools used in this thesis. We present a brief description for all of them and pause to explain the underlying principles of those central to our experiments.

3.1 Parsers and taggers

3.1.1 Constituency and dependency parsers

Many of the features extracted for several of our experiments are based on dependency parsing. We used the dependency parser of Henderson et al. (2008) and that of Bohnet et al. (2013) from the Mate toolkit. The first is a joint generative model of syntactic and semantic dependencies. The second is also a joint model which performs full morphological disambiguation and labeled dependency parsing. We chose these particular parsers based on their robustness, readiness and the availability of pre-trained models. Concerning the Mate parser, the models used in this thesis can be found online at `https://code.google.com/p/mate-tools/downloads/list`.

We have also used the rule-based constituent parser Fips (Wehrli 2007) as the base for an anaphora resolution method. A detailed description is provided in section 5.2.

3.1.2 Morphological analyzers

We used the Morfette system (Chrupała, Dinu, and van Genabith 2008), which produces joint morphological tagging and lemmatization. Morfette fits two separate logistic regression models: one for morphological tagging and one for lemmatization. The predictions of the models are then combined to produce a globally plausible sequence tag-lemma for the words of a sentence.

Morphological tags are more specific than the POS-tags produced by the parsers, allowing us to exploit them to create features for several experiments. Another advantage of these tags is that they are produced in context, then each form is associated with one tag, avoiding the ambiguity of other resources such as lexicons.

For some experiments where morphological tags were not needed, we have preferred the TreeTagger lemmatizer (Schmid 1994). TreeTagger produces POS-tags and lemmas exclusively, making it a very fast tool to parse large amounts of data. It models the probability of a tagged sequence of words (under a Markov assumption) using a binary decision tree to estimates transition probabilities.

Using pre-trained models in both cases, the Morfette system was employed for French, while we processed English and German with TreeTagger.

3.1.3 NADA

The Non-Anaphoric Detection Algorithm or NADA (Bergsma and Yarowsky 2011) is a system which distinguishes between the referential and non-referential English pronoun ‘it’. In this system, referential ‘it’ include nominal anaphoric instances only, while non-referential ‘it’ includes pleonastic instances and cases where the ‘it’ refers to a discourse segment. The system assigns probabilities to each instance of ‘it’, in a way that the higher probabilities represent a higher likelihood for ‘it’ to be referential. The system is a regularized logistic regression fed with lexical features indicating the presence of a particular string, and web count features taken from an auxiliary web-scale N-gram corpus.

3.2 Maximum Entropy classifiers

The maximum entropy model, MaxEnt, (Berger, V. J. Della Pietra, and S. A. Della Pietra 1996) and its implementation in the Stanford Classifier package (Manning and Klein 2003) is a recurrent classification instrument in this thesis. The maximum entropy model is a discriminative or conditional model which maximizes the entropy H of a conditional probabilistic model P(Y|X) that is trained to predict classes y ∈ Y for given data instances x ∈ X based on maximum likelihood estimation and parametrization of X and Y. The entropy can be understood as a measure of uniformity. The maximum entropy model is the model, in any set of models C (3.1 - 3.2).

$$\text{find } P_\star = \arg\max_{P \in C} H(P) \tag{3.1}$$

$$H(P) = \sum_x P_x \log\left(\frac{1}{P_x}\right) \tag{3.2}$$

The uniformity of the model, however, is subject to constraints by the data, i.e., the features. The construction of the model starts by defining features which are distinctive enough of the classes to differentiate. A feature f_i is a piece of evidence linking an observation x with a class y. The empirical counts of a feature $f_i(y, x)$ in the data are gathered (3.3) and used to compute the expected value of that feature f_i with respect to the observed distribution $P(y, x)$ (3.4).

$$E(f_i) = \sum_{(y,x) \in observed(Y,X)} f_i(y, x) \tag{3.3}$$

$$E(f_i) = \sum_{(y,x) \in (Y,X)} P(y, x) f_i(y, x) \tag{3.4}$$

Through the method of Lagrange multipliers, each feature f_i gets assigned a weight or parameter λ_i. The classes are then associated with the linear combination of the weights, $\sum \lambda_i f_i(y, x)$. The linear combination is used to produce the probabilistic model in (3.5), where $\tilde{y}$ represents the empirical distribution (as opposed to the predicted distribution y).

$$P(y|x,\lambda) = \frac{\exp \sum_i \lambda_i f_i(y,x)}{\sum_{\tilde{y}} \exp \sum_i \lambda_i f_i(\tilde{y},x)} \tag{3.5}$$

For the computation details in the Stanford package, the reader is referred to the system description provided by Manning and Klein (2003).

For the experiments reported in this thesis, the MAXENT model presented some characteristics which made it advantageous. First, it is a classifier which works well with training data of both relatively small size and large size as well. Second, it produces probability distributions over classifications, an aspect which allows to experiment with system combinations easily. Last, this model is not sensitive to feature overlapping, features are counted only once.

3.3 Statistical phrase-based machine translation

Like other NLP applications, POS-tagging or parsing, for instance, phrase-based machine translation processes texts one sentence at the time. Each sentence is taken as a vector of features and the translation model results from the combination of feature functions that have been trained independently. To train these functions, the sentence is decomposed into a set of phrases, hence the name *phrase-based* model. This type of phrases is typically composed by any consecutive sequence of words without prior knowledge about linguistic structure and without connection to the linguistic construct of phrase. The specific segmentation of a sentence into phrases depends on consistent word-level alignments, i.e., all words in a source phrase have alignment points in the corresponding target phrase and vice versa. This makes clear the importance of good word-alignment on top of which the phrase translation table is built Koehn (2010).

For any given sequence f, the translation model has the objective of finding the best translation e (3.6). The standard phrase-based model is the product of the combination of three main feature functions: a phrase table $\phi(\bar{f}|\bar{e})$ which ensures that the source phrases matches the target phrases, a reordering model d which accounts for the appropriate reordering of phrases, and a language model $p_{\mathrm{LM}}(e)$ responsible for ensuring fluent output.

$$\text{find} \quad e_{\text{best}} = \arg\max_{e} p(e|f) \tag{3.6}$$

$$e_{\text{best}} = \arg\max_{e} \underbrace{\prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)^{\lambda\phi}}_{\text{phrase table}} \underbrace{d(\text{start}_i - \text{end}_{i-1} - 1)^{\lambda d}}_{\text{reordering}} \underbrace{\prod_{i=1}^{|e|} p_{\text{LM}}(e_i|e_1 \dots e_{i-1})^{\lambda LM}}_{\text{language model}} \tag{3.7}$$

Note that these three components are assumed to be independent of each other, therefore they are independently trained. Then, there is an optimization step where each of these three components is assigned a weight λ (3.7), in order to give each one a difference in the value of their vote towards the final translation. The original probability $p(e|f)$ is inverted in (3.7) into $p(\bar{f}_i|\bar{e}_i)p_{\text{LM}}(e)$ through the Bayes' rule.

The model can be expressed in terms of a log-linear model by treating the components presented above as feature functions with weights attached to them. Each target sentence is then associated with the log-linear combination of the weights (3.8) of each feature function $h_i(f, e)$, in the exact same manner as the MAXENT model explained above. An inherent property of log-linear models is their ability to include additional feature functions easily. Factored translation models (Koehn and Hoang 2007) exploit this property.

$$e_{\text{best}} = \sum_{i=1}^{n} \lambda_i h_i(f, e) \tag{3.8}$$

Factored models integrate additional information into the model at the word level using supplementary mark up or *factors*. Other than the phrase table, the reordering model and the LM, factored models treat the added factors as a feature function $h_i(f, e)$ into the linear model in (3.8).

(1) Chomsky ran for an hour.
Chomsky|NULL ran|PC for|NULL an|NULL hour|NULL .|NULL

Any type of word-level information can be added as factors, and there can be several factors per word such as lemmas, POS tags, morphological tags, etc. In (1), we present an example from one of our experiments. We have experimented with a single disam-

biguating target side factor of the English simple past verb, in this case, the French tense. In the example, the simple past verb in the sentence receive one tense label, e.g. 'ran|PC' for *passé composé*, while all other words are set to the '|NULL' factor. The factor function is estimated in the same manner as the phrase translation table, based on consistent word-alignments points within the phrases (Figure 3.1).

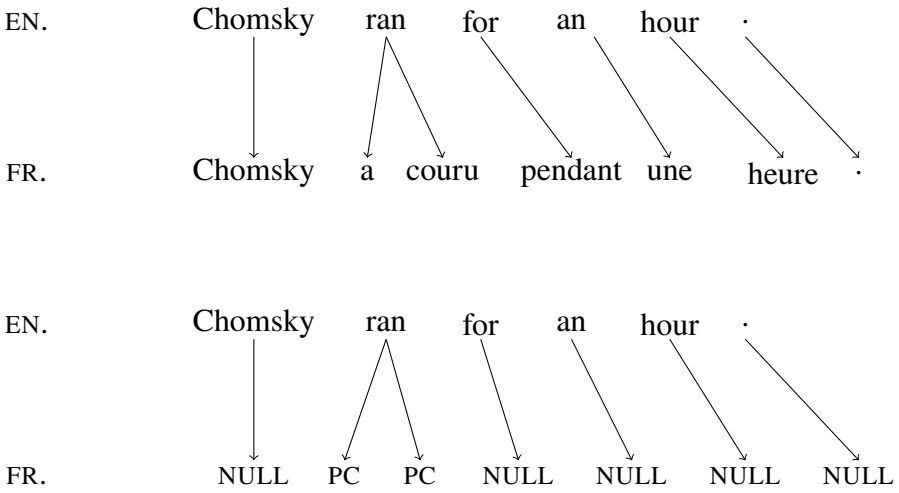


Figure 3.1: Example of word-level alignment for phrase extraction and simultaneous word-level factor alignment.

We have based our account of phrase-based models on Koehn (2010). We have used the Moses Toolkit (Koehn, Hoang, et al. 2007) to build our SMT systems, relying on GIZA++ (Och and Ney 2003) for word alignments and MERT (Och 2003) to optimize the weights of feature functions. Language models have been built using both SRILM (Stolcke et al. 2011) and KenLM (Heafield 2011).

Chapter 4

Pronouns across corpora and languages

4.1 Introduction

Pronouns are a prominent type of *anaphors*, elements without lexical signification which find their meaning by referring to another element with lexical content. This referring process is known as coreference between a pronoun and its *antecedent*. Although pronouns are not content words, as a result of the coreference process, they play a role in text understanding. Their function is to make communication easier and efficient, freeing the speaker of burdensome repetitions and constant explicitation of the referent or the antecedent (Webber 1979; De Beaugrande and Dressler 1981; Moeschler and Reboul 1994).

The complexity of the coreference relationship varies depending on the complexity of the antecedent the pronoun refers to, and the distance to which they stand from one another. As indicated by their name, antecedents of nominal anaphors are typically noun phrases (NPs). However, antecedents can also be long segments of text, including even entire clauses, as illustrated by the underlined *it* pronoun in example (1). These are the type of antecedents of eventual or abstract anaphors. They are much harder to identify and less extensively treated than their nominal counterparts (Dipper and Zinsmeister 2010; Dipper, Seiss, and Zinsmeister 2012; Kolhatkar 2015; Guillou 2016).

(1) Or I would decide I should have lunch, and then I would think, but I'd have to get

> the food out and put it on a plate and cut it up and chew it and swallow it, and *it* felt to me like the Stations of the Cross.

The antecedent carries its name because of its position in discourse. Indeed, most of the time the antecedent precedes the anaphor. The opposite positioning is also possible, in which case the phenomenon is called *cataphora*; however, this configuration is rare, reaching only 0.28% of the cases according to Laurent (2001). Regardless of their relative positioning from each other, if the antecedent is within the same sentence as the anaphor, then intra-sentential coreference occurs. Otherwise, when they are in two different sentences then inter-sentential coreference occurs.

Pronouns have morphological features involving their function, person, case, number and gender. Not all languages coincide in their use of these categories, resulting in translation mismatches susceptible to the pair of languages involved.

4.1.1 Types of pronouns

The categorization criteria of pronouns is debatable among linguists (Bhat 2004). However, this thesis focuses on a small number of languages and on a small set of pronouns. A central category for these languages –English, French and Spanish– is that of grammatical person. In this sense, *personal pronouns* distinguish particular grammatical persons, for instance, the subject pronouns *I*, *we*, *you*, *he*, *she*, *it*, *they* distinguish between the 1st, 2nd and 3rd person in English.

When the subject and the object of a sentence refer to the same entity, typically *reflexive* pronouns are used, e.g., *myself*, *ourselves*, *yourself*, *himself*. In Romance languages, reflexive pronouns are often *clitics*. Clitic is a term used for describing elements which are independent in meaning but unstressed and phonologically dependent tokens at the same time. For example, the English possessive *'s* is considered a clitic. These forms are particularly important in languages such as Spanish and Italian partly because they do not typically present overt subject pronouns, null subjects are preferred instead.

Null subjects, the omission of subject argument pronouns as in example (2), occur because the grammatical person features are inferred from the verb (Neeleman and Szendői 2005; Neeleman and Szendői 2007). Within linguistics, this characteristic is known as pro-drop, since an invisible pronoun *pro* is assumed to occupy the subject position. The resolution of this kind of subject is known as zero anaphora resolution (Ferrández and Peral 2000; Mitkov 2002; Mitkov 2003; Rello and Ilisei 2009).

(2) a. ES. Ø Escudriñ*a* todo lo que Ø tien*en*, y Ø encuentr*a* materiales para hacer su trabajo.
b. EN. And *he* digs through everything *they* have, and *he* finds materials to make work.

In pro-drop languages, an explicit pronoun is used mostly for stressing the subject, since mentioning the pronoun in every subject position results in an output perceived as less fluent (Clements 2008). The exception to this usage are impersonal sentences, in which the presence of a subject pronoun is not optional, it is ungrammatical. Pro-drop languages do not have expletive or pleonastic pronouns such as *it* or *there* in English, or *il* in French. The correct identification of expletive pronouns is crucial for coreference

resolution systems given that they are not genuinely referential pronouns and therefore do not have an antecedent. Expletive pronouns are associated with some categories of verbs, for instance meteorological verbs such as *to rain* or *to snow*. They are associated with some syntactic constructions as well, for instance, with adjectives which can take so-called sentential subjects when the sentence is extraposed as in (3a).

(3) a. ES. Ø Estaba claro que la guerra ocurriría.
b. EN. *It* was clear that war would happen.

Other types of pronouns include indefinites, possessive, demonstrative and relative pronouns. Indefinites (*everyone*, *someone*, etc.) are autonomous expressions that do not corefer (within the text) (Zufferey and Moeschler 2012). Possessive pronouns are used to refer to two elements at the same time: a possessor and a possessed thing. Halliday and R. Hasan (1976, p. 45) illustrate this relationship with the example *'Can you hand Mary a program? Hers has got lost'*, in which *hers* sends back to both *Mary* and *programme*. Demonstrative pronouns are used when referring to a location on a scale of proximity: *this*, *these* opposed to *that*, *those* (Halliday and R. Hasan 1976).

4.2 Pronouns across corpora and languages

The distribution of pronouns differs from one type of corpus to another. To illustrate this, we have looked into a literary corpus[1] and a journalistic corpus [2]. Both corpora were tagged using the Fips parser (Wehrli 2007), and the categories of the pronouns considered are those produced by the parser. A small sample of 500 sentences shows that the first contains substantially more pronouns (910 vs 259 for English). The distribution across languages is also different as shown in Figure 4.1. Italian for example, has fewer personal pronouns than English, French and German because of the pro-drop instances. Inversely, the number of personal pronouns is higher in English, French and German because there are no pro-drops in these languages. Finally, Italian has more clitic pronouns than French because verbal pronominal phrases are preferred where in French the passive voice is used, another consequence of the pro-drop feature (Scherrer

[1] *The Little Prince* by Antoine de Saint-Exupéry. The French, English and German versions were downloaded from `http://wikilivres.info`, and the Italian one from `http://www.macchianera.net/files/ilpiccoloprincipe.pdf`

[2] 2007 press releases, as available at: `http://www.news.admin.ch`

et al. 2011).

A closer description of English and Spanish personal pronouns is presented in Table 4.1 and Table 4.2 respectively. We use the NewsTest2013 corpus from the WMT 2013 data (Bojar, Buck, Callison-Burch, et al. 2013) for these figures. To identify the pronouns, the English corpus is parsed with the dependency parser of Henderson et al. (2008) and the POS-tags are used. The Spanish corpus is analysed with Morfette, which produces morphological tags and lemmas. The quality of these tools is very good in general, however, some degree of error is to be expected coming from the ambiguity of some pronoun forms. For instance, there is an ambiguity between demonstrative determiners and demonstrative pronouns in examples such as *This light gives me headache* vs *This is better*.

For English, 4,864 pronouns were found, 8.67% of the total words in the corpus. The Spanish corpus, on the other hand, has 3,866 pronouns or 6.23% of the total words in the corpus. From this, 43.3% are personal pronouns in English and 38.4% in Spanish. Although the total amount of pronouns is comparable, their distribution is very different. While in English (regular) personal pronouns outnumber reflexives (97.92% vs 2.08%), the relationship is less extreme in Spanish, (45,27% vs 54.73%). Spanish reflexives often appear substituting the direct and indirect objects and often acting as pseudo-subjects, as is the case of the reflexive pronoun *se* in example (4) (Pineda and Meza 2006).

(4) a. ES. *Se* sospecha que *se* ha embolsado repetidos sobornos a cambio de contratos públicos.

b. EN. *He* is suspected of pocketing repeated bribes in exchange for public contracts.

In the rest of this work, we will focus on personal pronouns mainly. This a difficult type of pronoun to translate in the context of MT due to the distance they can take from their antecedent, the complexity of the antecedent itself and the particular usages of the languages involved.

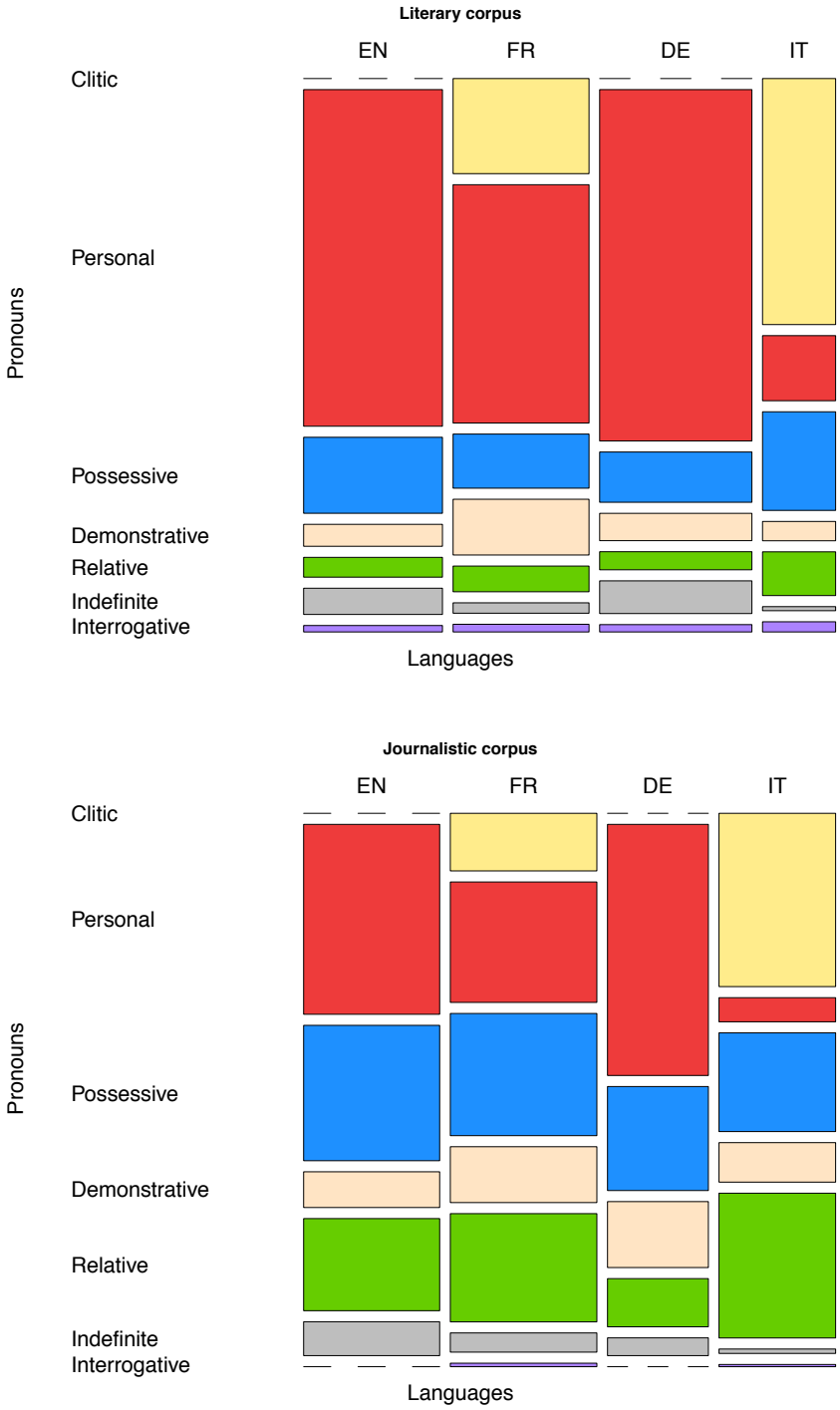


Figure 4.1: Comparison of of the distribution of different types of pronouns in 500 sentences from the Little Prince and from Press Releases.

Type	Pronoun	#	%
Personal	I	367	16.64
	you	200	9.07
	they	245	11.11
	she	39	1.77
	he	306	13.88
	it	514	23.31
	we	232	10.52
	us	56	2.54
	me	52	2.36
	her	2	0.1
	him	61	2.77
	them	84	3.81
Reflexive	itself	12	0.54
	ourselves	5	0.22
	himself	11	0.5
	myself	2	0.1
	themselves	16	0.72
Total		2,204	100

Table 4.1: Distribution of personal pronouns in the English NewsTest2013 corpus.

4.3 Word-level alignment and pronoun translation mapping

Word-level alignment is an essential part of phrase-based statistical machine translation (SMT). This process builds on bilingual corpora aligned at the sentence-level. The process ends when all words in the source sentence are paired with their corresponding translation. Not all alignments are one-to-one. A word in either side of the corpus might be aligned to several words or to no word at all in the other side, reflecting the mismatches in linguistic structures across the languages. For instance, function words which exist in one language but not the other and idioms are known problematic cases for word-level alignment (Koehn 2010). Figure 4.2 illustrates this process with the French sequence *de l'* which gets aligned to *with* and the preposition *de* which is assigned the artificial NULL alignment.

Some of the difficulties of pronoun translation are reflected in these quirks of word-alignment. Besides, the alignment process can be exploited to quantify and understand how pronouns are translated. For this purpose, we have built English-French word-

Type	Pronoun	#	%
Personal	consigo	1	0.07
	él	40	2.69
	ella	17	1.14
	ellas	13	0.87
	ello	24	1.61
	ellos	48	3.22
	la	7	0.47
	las	1	0.07
	le	101	6.78
	les	59	3.96
	lo	100	6.72
	los	4	0.27
	me	81	5.44
	mí	12	0.81
	nos	54	3.63
	nosotros	20	1.34
	sí	15	1.01
	ti	4	0.27
	usted	12	0.81
	vosotros	1	0.07
	yo	36	2.42
Reflexive	te	1	0.07
	se	785	52.72
	nos	9	0.60
	me	20	1.34
Total		1489	100

Table 4.2: Distribution of personal pronouns in the Spanish NewsTest2013 corpus.

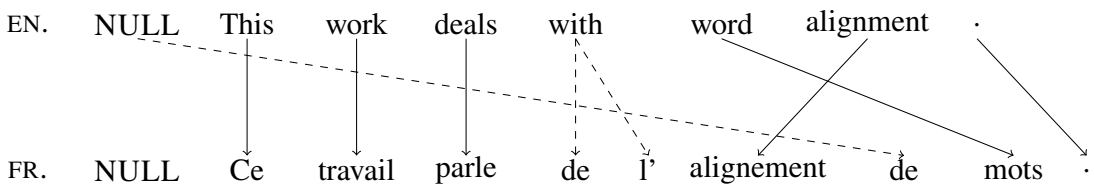


Figure 4.2: Example of word-level alignment correspondences.

level alignments using GIZA++ and the NewsTest2013 corpus from the Eight Workshop on Statistical Machine Translation (WMT 2013) translation task data (Bojar, Buck, Callison-Burch, et al. 2013). In order to obtain sufficient statistical evidence for precise alignments the NewsTest2013 data was concatenated with the Europarl version 7 corpus provided as training data. Table 4.3 and Table 4.4 contain all translation possibilities as found in the corpus. In particular, we noted that pronouns are not necessarily translated by a single pronoun, but many of them correspond to some reformulation without a pronoun (NULL alignment) as in example (5), or to some lexical translation, collected under the catch-all category OTHER.

(5) a. EN. The recognition is going to be slow in the United States, no question about that, but in the UK, it is happening, and in other countries *it* is happening.

b. FR. Cette reconnaissance est lente au USA, c'est certain, mais en Angleterre, c'est en train de se passer, et _ aussi dans d'autres pays.

A look at Table 4.4 quickly reveals the diversity of translation a pronoun may have. Take for instance the pronoun *it*. This pronoun is not only the most frequent, but also may be translated with 24 different possibilities. Some of the translations reflect the source pronoun function. For example, *ce* and *cela* are likely translations of the pronouns with event reference function. In order to have a relative measure of the quality of pronoun translation by phrased-based systems, we used the WMT 2013 training data and the MOSES toolkit (Koehn, Hoang, et al. 2007) to build a phrase-based system. The system was tuned on the NewsTest2010 corpus using MERT. A 5-gram language model trained on the entire French side of Europarl version 7 with SRILM is used. A comparison of the source-candidate alignment with the source-reference alignment of the NewsTest2013 data revealed that only 38% (180/514) of the translations of the *it* pronouns were identical between the two. Furthermore, reflexive pronouns had very few matches as well (*ourselves* 2/5 matches, *himself* 3/11 matches, *themselves* 5/16 matches, *itself* 4/12 matches). These tendencies are in line with results reported by Hardmeier and Federico (2010) and Hardmeier (2014). These authors examined a set of 219 sentences and found that feminine singular pronouns, pronouns of polite address, reflexive pronouns and the combination pronoun+preposition, were the worst translated categories by German-English system. They also pointed out to the large size of the OTHER category.

TARGET↓	themselves	myself	ourselves	himself	itself	total
OTHER	2		1	1	2	6
NULL	2			2	6	10
eux-mêmes	3					3
eux	1					1
leur	1					1
se	5		1	5	3	14
ils	2					2
me		2				2
nous			1			1
nous-mêmes			1			1
notre			1			1
lui-même				3	1	4
total	16	2	5	11	12	46

Table 4.3: Target pronoun translation of English reflexives pronouns.

4.3.1 Translation into French of English *it* and *they*

In the context of the 2015 DiscoMT shared-task on cross-lingual pronoun prediction (Hardmeier, Nakov, et al. 2015), nine classes of French pronoun were defined as possible translations for the English personal pronouns *it* and *they*. We determined their distribution relying on the word alignments provided with the training data and which we corrected by hand. Specifically, 446 instances of pronouns aligned to random words were corrected by hand. Their distribution is presented in Table 4.5 and Figure 4.3. The important imbalance of the OTHER class in particular is to be noted. This category stands for cases where the translation corresponds to something which is not a pronoun and it amounts to $\approx 20\%$ of the translations.

A manual evaluation of the data confirmed that the OTHER class includes translations as lexical NPs (6) and other pronouns not defined for the task (7). Object pronouns are included as well (8). The NONE class, on the other hand, corresponds to English pronouns which were not translated at all in French and which are assigned the NULL alignment, as in the case of paraphrases (9). Similar proportions were reported by Weiner (2014) for the translation from English to German.

(6) a. Certainly *it* is perceived de facto to be impossible.
b. *La chose* est certainement perçue de facto comme étant impossible.

TARGET↓	I	you	they	she	he	it	we	us	me	her	him	them	**total**
OTHER	10	5	14	1	6	17	3	15		1	1	9	82
NULL	24	50	47	7	63	125	15	10	6		15	28	390
me	8				1				32				41
on	3	18	9		2	28	40						100
ma	1												1
je	309					4			2				315
elle	1		2	28		44							75
moi	3								10				13
mon	2								1				3
mes	1								1				2
il	5	6	11	1	209	123	10				1		366
vous		101											101
tu		12											12
votre		1											1
toi		2											2
vos		2											2
nous		3	2			3	159	31					198
elles			12			1						1	14
eux			3									8	11
leur			3									11	14
ils			129			3	1						133
qui			1			2							3
se			1		1	3							5
ce			3		6	83							92
ça			1			11							12
cela			2			31							33
ces			4			1						2	7
ceux-ci			1										1
ses				1	2								3
lui				1	8	4					35		48
celui					1								1
son					2	3				1	1		7
lui-même					1								1
que					3	1							4
sa					1						1		2
en						1							1
le						16	2				7		25
la						3							3
celui-ci						1							1
cette						4							4
y						2							2
nôtre							1						1
nos							1						1
les												25	25
total	367	200	245	39	306	514	232	56	52	2	61	84	2158

Table 4.4: French pronoun translation of 2,158 English personal pronouns from the NewsTest2013 data.

French Translation	**it**		**they**	
	#	%	#	%
ça	79	0.43	1	0.02
cela	585	3.19	22	0.33
elle	2,392	13.03	93	1.40
il	5,332	29.04	275	4.14
ce	1,919	10.45	128	1.93
elles	101	0.55	911	13.72
ils	158	0.86	3,263	49.13
on	360	1.96	97	1.46
NONE	2,895	15.77	515	7.75
OTHER	4,537	24.71	1,337	20.13
Total	18,358	100.00	6,642	100.00

Table 4.5: Distribution of the French Translations of English pronouns *it* and *they* in 25,000 occurrences from TED talks (747 examples), News Commentary (14,561 examples) and EuroParl version 7 (9,691 examples).

(7) a. It [a budget line] was not able to do very much but *it* was repeatedly abused by Members of this House proposing action when the disasters were not even major.

b. Elle [une ligne budgétaire] ne permettait pas de faire grand-chose mais les députés de cette Assemblée *en* abusaient constamment en proposant d'agir alors que l'ampleur des désastres n'était même pas importante.

(8) a. We have that opportunity right now. Let us grasp *it*.

b. Cette chance se présente aujourd'hui, et nous devons *la* saisir !

(9) a. I believe *it* to be of vital importance that where Member States allow regions and local authorities to raise taxes, *they* should continue to be able to do so and not be subject to across-the-board regulation by Europe.

b. Je voudrais dire que j'estime indispensable que les États membres puissent continuer d'autoriser les régions et les communes à percevoir des taxes et que ce domaine ne soit pas uniformément réglé par l'Europe.

These figures and examples show that pronoun translation is not deterministic, it depends on the context and the preferences of the target language. For instance, many *on* occurrences are translations of English passive constructions of verbs with two objects

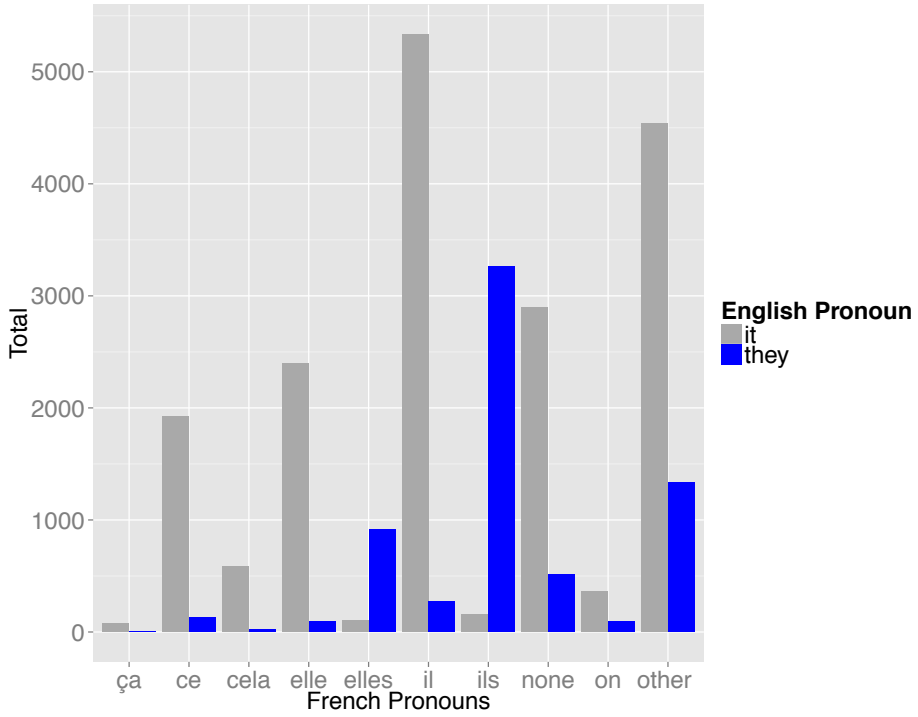


Figure 4.3: Distribution of the French Translations of English pronouns *it* and *they* in 25,000 examples sampled from TED talks (747 examples), News Commentary (14,561 examples) and EuroParl version 7 (9,691 examples).

such as *to give* in (10). The promotion of a second object to subject is impossible in French and an active construction is preferred instead.

(10) a. EN. And gamers are willing to work hard all the time, if *they're given* the right work.

b. FR. Les joueurs sont prêts à travailler dur, tout le temps, si l'*on* leur confie la bonne mission.

Moreover, there is no equal distribution of the genders in French. The masculine pronouns *il* and *ils* are non-marked classes used for masculine antecedents, impersonal uses, and plural antecedents referring to both genders. Therefore, there is a strong presence of these translations in contrast to the feminine forms.

4.4 The problem of pronoun translation evaluation

Automatic MT metrics are based on similarity measures between the reference translation and the system output. Their purpose is to set a crossbar which allows measuring output quality of a system. In contrast to manual evaluations, it is evident that automatic metrics are easy to implement, and a fast and cheap way to estimate the quality of automatic translation. Likewise, these metrics are needed during training and tuning of the MT algorithms in order to know when a system is already calibrated for good quality machine translation. These measures, however, are subject to debate and their many shortcomings have been pointed out, e.g., decreased performance with a single reference, lack of qualitative criteria such as lexical choice, unsatisfactory account of recall (Koehn 2010). Other critics point to the fact that not all mistranslations have the same degree of unacceptability, something an automatic metric cannot grasp (Song, Cohn, and Specia 2013).

The lack of an appropriate metric to measure the quality of pronoun translation has been discussed in practically all papers dealing with the subject of pronoun translation or integrating discourse level phenomena into MT. Automatic metrics are not sensitive enough to grasp any change in a single linguistic phenomenon such as pronouns. At the same time, human evaluation is not efficient (Webber 2014).

In this context, Hardmeier and Federico (2010) suggested a metric for evaluating pronoun translation based on the BLEU (Papineni et al. 2002) clipped counts. BLEU stands for *Bilingual Evaluation Understudy*, and it is by far the most used metric to evaluate MT output. It is based on a precision measure computed between one or multiple human reference translations and the candidate translation generated by a system. It involves three main components: a) n-gram precision computed on the basis of b) clipped counts and c) a brevity penalty. A clipped count would be the number of times a word occurs in the hypothesis, limited by the number of times it occurs in the reference (4.1). It ensures that over-generation or repetition of words do not inflate the score. The n-gram precision measures how well a candidate translation matches the reference translation(s) (4.2). Finally, the brevity penalty decreases with candidates which are shorter than the reference(s) (4.3). The BLEU metric is given in (4.4), where w is the weight of each reference n, usually $w_n = 1/\mathrm{N}$.

$$c_{clip}(w) = \min(c_{\mathrm{C}}(w), c_{\mathrm{R}}(w)) \tag{4.1}$$

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')} \tag{4.2}$$

$$\text{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{(1-r/c)} & \text{if} \quad c \leq r \end{cases} \tag{4.3}$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.4}$$

Hardmeier and Federico (2010)'s metric basically consists in counting the correct pronoun translations and in computing precision and recall based on the clipped counts of pronouns. A translation is considered correct when the pronouns in the source-reference and source-hypothesis alignments match. We reproduce their formulas below (c stands for count, C stands for candidate translation, R is the reference translation, and last, *w* stands for word, in this case a pronoun).

$$\text{Precision} = \frac{\sum_{w \in \text{C}} c_{clip}(w)}{|\text{C}|}; \text{Recall} = \frac{\sum_{w \in \text{C}} c_{clip}(w)}{|\text{R}|} \tag{4.5}$$

As pointed out by Webber (2014) mismatches in the source, candidate and reference translations vary a lot depending on the languages one is dealing with. For instance, there might be a pronoun in the source and the automatic translation, but not in the reference. The clipped counts in Hardmeier and Federico (2010)'s metric take this into account but do not entirely account for the uncertainty in the translations which are not pronouns.

Weiner (2014) addresses the problem of just comparing the hypothesis with the reference and proposes to evaluate the anaphora resolution itself. Using a fine-grained POS tagger, he evaluates (using precision and recall) if the target language antecedent and pronoun agree in their gender and number features. This strategy involves a coreference resolution system in the target language which together with the POS tagger performance introduce many errors.

Meteor (Denkowski and Lavie 2011) is another existing automatic metric to score MT output. Since we also used it in some of our experiments, we describe it here for completeness. This score is also based on a similarity measure between a candidate (h) and a reference (r) translation. However, the matching between the two is more elaborate than the n-gram matching on which BLEU is built. Meteor relies on the alignment of matching tokens, stems, synonyms and paraphrases (which are the matchers m_i). With these counts, the harmonic mean of the precision and recall is computed (equations (4.6), (4.7) and (4.8)). Besides, the score distinguishes between function (f) and content words (c) and a function-content word weight (δ). The Meteor score also accounts for word order discrepancies by means of a fragmentation penalty (4.9) which is computed using the total number or matched words and the number of matched chunks (continuous word matches). The final score is presented in (4.10). The parameters α, β, γ, δ and $w_i...w_n$ need to be tuned as well.

$$\text{Precision} = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|} \tag{4.6}$$

$$\text{Recall} = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|} \tag{4.7}$$

$$F_{\text{mean}} = \frac{\text{Precision} \cdot \text{Recall}}{\alpha \cdot \text{Precision} + (1-\alpha) \cdot \text{Recall}} \tag{4.8}$$

$$\text{Penalty} = \gamma \cdot \left(\frac{\text{chunks}}{\text{matched words}} \right)^{\beta} \tag{4.9}$$

$$\text{Meteor} = (1 - \text{Penalty}) \cdot F_{\text{mean}} \tag{4.10}$$

For our experiments with machine translation systems, we report the obtained automatic metrics of output quality. However, since they are not entirely suitable for targeted aspects of a translation, we also report the results of comprenhensive manual evaluations. Our cross-lingual pronoun prediction experiments, on the other hand, are evaluated as standard classification tasks using a gold standard test set.

4.5 Conclusion

The purpose of this chapter has been to give a working description of the type of pronouns treated in subsequent chapters. Besides, it has been shown that the different types of pronouns have different distributions across corpora and languages. It has also been shown that the translation of each pronoun presents a multiple choice depending on the particular preferences of the target language. Finally, we have described two widely used MT metrics and another one specific to the problem of pronoun translation.

Chapter 5

Pronoun translation

5.1 Introduction

The interest for pronoun translation is at the heart of a line of research concerned with discourse phenomena and Machine Translation (MT). Many translation problems concern long distance dependencies and phenomena beyond the sentence level. Research focused on the discursive properties of texts has been active for years and its insights can be now employed to improved MT.

The linking, or resolution, of a pronoun and its *antecedent* seems trivial for a human, but is not straightforward for a machine, especially if the antecedent and the anaphor are not in the same sentence and the text in question contains several sentences with several potential antecedents. Developing automatic Anaphora Resolution (AR) systems is a research domain on its own and has been active for decades (cf. Section 2.2.2).

(1) a. Paul left two *bikes* in front of the house. When he came back, *they* were no longer there.

In addition, if sentence (1), for instance, is to be translated into French, one has the choice (mainly) between *ils* and *elles* for translating the pronoun *they*. This choice is no longer dependent on the English antecedent 'bikes', but on its translation in French either as the masculine noun *vélos* (2a) or as the feminine noun *bicyclettes* (2b).

(2) a. Paul a laissé deux *vélos* devant la maison. Lorsqu'il est revenu, *ils* n'étaient

plus là.

b. Paul a laissé deux *bicyclettes* devant la maison. Lorsqu'il est revenu, *elles* n'étaient plus là.

The focus of this chapter is on the English third person pronouns *it* and *they* and their translation into French. As observed in corpora (Section 4.3.1), these pronouns are not always translated as pronouns, but can correspond to a content noun phrase (NP) or to nothing at all. This is the case in example (3) where the English pronoun *they* in (3a) corresponds to a content NP in French (3b).

(3) a. To conclude, I would just like to say something on the principle of subsidiarity. I believe it to be of vital importance that where Member States allow regions and local authorities to raise taxes, *they* should continue to be able to do so and not be subject to across-the-board regulation by Europe.

b. Enfin, concernant le principe de subsidiarité, je voudrais dire que j'estime indispensable que *les États membres* puissent continuer d'autoriser les régions et les communes à percevoir des taxes et que ce domaine ne soit pas uniformément réglé par l'Europe .

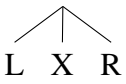
In this chapter, we first address the problem of pronoun translation as a classic *Anaphora Resolution* (AR) task before translation. Our AR procedure is then applied to a Rule-Based Machine Translation (RBMT) context. Unlike Statistical Machine Translation (SMT) systems, this framework allows to tackle the translation of null subjects explicitly. In this framework, there are works which use an AR component before translation time as well. These efforts, however, have achieved only modest results (cf. Section 2.2.3).

In the second part of this chapter, we follow the line of work initiated by Hardmeier (2014) which have formalized pronoun translation as a cross-lingual pronoun prediction task (cf. Section 2.2.3). The cross-lingual pronoun prediction systems described in this chapter are not developed nor intended as AR systems. Therefore, they do not explicitly search the antecedent of pronouns, but their purpose is to predict directly a pronoun translation using a classifier fed with features extracted from parallel data. They represent an alternative to the use of an AR system for helping MT. In our experiments, we test the relevance of syntactic, morphological and contextual features for this task.

5.2 Rule-based machine translation of pronouns

Its-2 (Wehrli and Nerima 2009; Wehrli, Nerima, and Scherrer 2009) is a rule-based translation system based on the Fips parser (Wehrli 2007). The translation process follows the three classic steps: analysis, transfer and generation. Start with the analysis module. For a given source language sentence, the parser produces an information-rich phrase-structure representation, along with predicate-argument labels. The grammar implemented in the Fips parser is heavily influenced by Chomsky's minimalism program and earlier work (Chomsky 1995), but also includes concepts from other theories such as LFG (Bresnan 2001) and Simpler Syntax (Culicover and Jackendoff 2005). The syntactic structures built by the parser follow the general X-bar schema shown in (4), which yields relatively flat structures, without intermediate nodes.

(4) $[_{\text{XP}}$ L X R] XP



Each constituent XP is composed of a head, X, along with a (possibly empty) list of left sub-constituents (L) and a (possibly empty) list of right sub-constituents (R), where X stands for the usual lexical categories – N(oun), V(erb), A(djective), Adv(erb), P(reposition), C(onjunction), etc., to which we add T(ense) and F(unctional). The T category stands for tensed phrases, corresponding, roughly, to the traditional S category of standard generative linguistics. As for F, it is used to represent secondary predicates, as in the so-called small clause constructions.

The transfer module maps this source language abstract representation to an equivalent target language representation. The mapping is achieved by a recursive traversal of the source-language structure, starting with the head of a constituent, and then its right and left subconstituents. Lexical transfer occurs at the head level and yields a target language equivalent term of the same or different category, which becomes the new current head. The target language structure is then projected on the basis of the head. In this way, the final output is generated according to the lexical features of the target language. Argument constituents, on the other hand, are determined by the subcategorization properties of the target language predicate. The necessary information is available in the lexical database. Transformational rules, in the traditional Chomskyan sense, can apply to generate specific structures such as passive or *wh*-constructions (interrogative, relative,

tough-movement[1]). In addition, the transfer procedure can be augmented with language-pair specific transfer rules, for instance to modify the constituent order.

Currently, the Its-2 system is available for ten language pairs between English, French, German, Italian and Spanish. For each language pair, there is a bilingual, bidirectional dictionary implemented as a relational table containing the associations between the lexical items of source and target languages. Other specifications such as translation context, semantic descriptors and argument matching for predicates are also contained in the table.

In the Its-2 system, pronouns are handled like other lexical heads, that is, they are transferred and translated as heads of phrases, using the bilingual dictionary. This strategy, which works fine for non-anaphoric pronouns, is clearly insufficient for anaphoric pronouns, for which knowledge of antecedent is mandatory. The following section describes the implementation of an anaphora resolution component in the Its-2 system, as part of the Fips parser. This AR component only deals with 3rd person personal pronouns such as (*he*, *she*, *it*, *her*, *him*, etc.). The basic idea underlying our implementation is that the proper form of a target-language pronoun depends on the gender and number features of its (target-language) antecedent. Since we do not perform AR on the target language, this information can be retrieved through the links connecting the source-language pronoun, its antecedent and the target-language correspondence of the antecedent. To illustrate this process, consider the following example:

(5) a. EN. Paul bought an ice-cream and will eat *it* later.
b. FR. Paul a acheté une glace et *la* mangera plus tard.

The pronoun *it* in the source language should be translated as a feminine (clitic) pronoun **la** in the French sentence, because *ice-cream*, the antecedent of *it*, is translated as *glace*, a feminine noun.

[1] *tough*-movement refers to subjects of a main verb which are also the object of an embedded infinitive verb. In '*This book is easy to read*', for instance, *this book* is both the subject of the main verb and the logical object of the verb *to read*.

5.2.1 Pronominal anaphora resolution based on Binding Theory

As indicated above, the AR procedure presented here is part of the Fips parser and it is conceived to deal with 3rd person personal pronouns. The AR procedure is highly influenced by Chomsky's Binding Theory Chomsky (1981), which is not an AR method per se, but rather a set of constraints useful to exclude otherwise potential antecedents. These constraints follow two principles: **Principle A** states that reflexive and reciprocal pronouns find their antecedents within their governing category (the smallest clause that includes them); **Principle B** states that 3rd person personal pronouns find their antecedents outside of the clause that includes them (Reinhart 1983; Büring 2005).[2]

Our strategy for anaphora resolution recalls in several ways the one used by Hobbs (1978) or Lappin and Leass (1994), adapted to the specific structures of the Fips parser.

The algorithm comprises three steps:

1. **impersonal pronouns**
   The impersonal pronoun *it* in English – *il* in French – has no antecedent and should be excluded from further consideration by the AR procedure. The identification of impersonal pronouns is achieved on the basis of lexical information (verbs lexically marked as impersonal, for instance meteorological verbs such as 'to rain' or 'to snow'), as well as syntactic information. For instance, adjectives which can take so-called sentential subjects occur with an impersonal subject when the sentence is extraposed as in:

   (6) a. *It* was obvious that Paul had lied.
   b. *It* is easy to see that.

   Similarly, impersonal subject pronouns can be found in passive structures with sentential complements:

   (7) *It* was suggested that Paul would do the job.

2. **reflexive or reciprocal pronouns**
   We assume a simplified interpretation of **Principle A** according to which a reflexive or reciprocal pronoun always refers to the subject of the smallest clause that

[2]Notice that Binding Theory includes a third principle, Principle C, which states that referring expressions (lexical noun phrases) cannot be bound. This principle is not relevant in this work.

contains it. In cases of embedded infinitive sentences, we assume the presence of an abstract subject pronoun (PRO, unrealized lexically) whose antecedent is determined by the *control theory* and ultimately by lexical information. For example, in the sentence *Paul$_i$ promised Mary [PRO to take care of himself$_i$]*, *himself* refers to the subject pronoun PRO, which in turn refers to the noun phrase *Paul*.

3. **referential non-reflexive/reciprocal pronouns**
   Such pronouns, currently restricted to the non-impersonal *it*, along with *he*, *him*, *she*, *her*, *they*, *them*, etc., undergo our simplified interpretation of Principle B, which states that they must have an antecedent outside of the clause that contains them. We further restrict possible antecedents to arguments, excluding adjuncts noun phrases. The search for antecedents considers all preceding clauses within the sentence as well as within the previous sentence and makes an ordered list of the noun phrases which agree in number and gender with the pronoun. The order is determined by proximity, as well as by the grammatical function of the antecedent (subject, then grammatical object, then prepositional complements, etc.).

In summary, the AR procedure is based on a simplified interpretation of the principles A and B of the Binding Theory. After attempting to eliminate impersonal pronouns, the procedure uses principles A and B, respectively to handle reflexive/reciprocal pronouns and other 3rd personal referential pronouns. Our simplified interpretation of those principles state that reflexive/reciprocal pronouns can only refer to the subject of their clause, while other pronouns can refer to noun phrases outside of their immediate clause. When several noun phrases meet those conditions, priority is given to grammatical function and locality.

Experiment 1: Pronoun translation with AR

The modified Its-2 system with the AR component was evaluated in the context of the shared task on pronoun translation organized by the DiscoMT 2015 Workshop (Loáiciga and Wehrli 2015; Hardmeier, Nakov, et al. 2015). The shared task focused on the translation of pronouns *it* and *they* into French. In order to ease the evaluation process, acceptable translations were limited to the following eight possibilities: *ce, elle, il, elles, ils, ça/cela, on* and OTHER.

Both the development and test sets provisded consisted in TED talks transcriptions (Hardmeier, Tiedemann, Nakov, et al. 2016). The test data was composed of 2,093 segments

including 1,107 *it* and *they* pronouns. For the official evaluation results, a sample of 210 pronouns from each submission and the baseline system provided by the organisers were judged by human evaluators. In order to deal with the variance in translation output across systems, annotators were not presented with reference translations. Instead, pronouns were removed from the translations and the annotators were asked to fill in the gaps, producing a pronoun which fits the rest of sentence and up to 5 sentences of context.

The official scores were computed comparing the systems' output with the manual annotations of the human judges. In the case that it were impossible to determine an adequate pronoun translation, the annotators had the possibility to annotate the segment as *bad translation*. For our system, there were 15 such cases. The results per pronoun translation obtained by our system are reported in Table 5.1.

Translated pronouns	**Precision**	**Recall**	**F1**
ce	0/ 0	0/ 41	n/a
cela	0/ 0	0/ 52	n/a
elle	5/ 19 (26.3%)	5/ 20 (25.0%)	25.6%
elles	3/ 6 (50.0%)	3/ 11 (27.3%)	35.3%
il	19/ 77 (24.7%)	19/ 25 (76.0%)	37.3%
ils	34/ 41 (82.9%)	34/ 46 (73.9%)	78.2%
on	0/ 0 (n/a)	0/ 0 (n/a)	n/a
OTHER	27/ 52 (51.9%)	27/ 30 (90.0%)	65.9%
All pronouns	61/143 (42.7%)	61/165 (37.0%)	39.6%
Accuracy with OTHER 88/210 = 0.419 Accuracy without OTHER 61/180 = 0.339			

Table 5.1: Results of the translation of 210 English pronouns into French using the RBMT system Its-2 with AR procedure.

The official shared task score was accuracy. Our system obtained an accuracy of 0.419 without translations as OTHER and 0.339 with OTHER. These results were rather low when compared with the other submitted systems. One obvious problem that can be observed is that our system only generates *il, elle, ils* and *elles* as possible translations of *it, they*. The non-generation of *ça, cela, ce* penalizes the results heavily. This is the case of example (8), where a translation of *it* as *ça* or *cela* would have been preferable. Yet, there is an effect of the AR component, visible in the generation of pronoun *elle*.

(8) a. SRC And when I was an adolescent, I thought that I'm gay, and so I probably

can't have a family. And when she said it, *it* made me anxious.

b. W/O AR Et quand j'étais un adolescent, j'ai pensé que je suis gai, probablement et ainsi je ne peux pas avoir une famille. Et quand elle l'a dit *il* m'a rendu anxieux.

c. W/ AR Et quand j'étais un adolescent, j'ai pensé que je suis gai, probablement et ainsi je ne peux pas avoir une famille. Et quand elle l'a dite *elle* m'a rendu anxieux.

To have a sense of the performance of the AR component in particular, we evaluated two translations of the test set, one using the AR component and another one without it. Unsurprisingly, the translation of the test set using the AR component does not have an impact on the BLEU scores. When measuring only the impact on pronoun translations, however, the AR component shows a positive effect in precision when compared to a baseline without it, as shown in Table 5.2. Since these results are computed using exact word-level alignment matching between the candidate translation and an unique reference (Hardmeier, Nakov, et al. 2015), they are only indicative.

	BLEU		**Precision**	**Recall**
with AR	22.43	it	0.1174	0.1173
		they	0.3631	0.3481
without AR	22.44	it	0.0917	0.0919
		they	0.2710	0.2566

Table 5.2: Contrastive results obtained from the translation of the test set with and without the AR component. Precision and recall scores were computed using the automatic score by Hardmeier and Federico (2010) (cf. Equation 4.5, Chapter 4).

In addition, we randomly selected 405 sentences with 203 pronouns and completed our own manual evaluation of two translations with and without the AR component. Results are summarized in Table 5.3. This manual evaluation focuses on the pronouns that the system is programmed to address.

It can be seen that the reflexive/reciprocal pronouns did not change between the two outputs. Besides, all observed errors were due to incorrect antecedent identification, leading to incorrect pronoun generation. One such a case is (9), where the algorithm turns a correctly translated pronoun by the baseline into an incorrect one. In this example, the word *procedures*, which is feminine in French, is identified as antecedent, causing then the generation of *elles* instead of *ils*.

EN Pronoun	Improved	Unchanged	Degraded
him	0	17	0
it	18	86	6
them	0	21	0
themselves	0	1	0
they	2	47	5
Total	20	172	11

Table 5.3: Results obtained from the manual evaluation of translation of 203 pronouns from the test set with and without the AR component.

(9) a. SRC And he spent all this time stuck in the hospital while he was having those *procedures*, as a result of which he now can walk. And while he was there, *they* sent tutors around to help him with his school work.

b. W/O AR Et il a passé tout ce temps englué dans l'hôpital tandis qu'il avait ces procédures, comme un résultat de lequel maintenant il peut marcher. Et tandis qu'il était là-bas, *ils* ont envoyé des professeurs autour pour l'aider avec son école à travailler.

c. W/ AR Et il a passé tout ce temps englué dans l'hôpital tandis qu'il avait ces *procédures*, comme un résultat de lequel maintenant il peut marcher. Et tandis qu'il était là-bas, *elles* ont envoyé des professeurs autour pour l'aider avec son école à travailler.

In almost the double of cases, however, the AR works in favor of a better pronoun translation. This is the case in example (10). Here the word *acceptance* is correctly identified as the antecedent. This translates as the feminine *acceptation* in French, therefore, the pronoun *it* is translated as *elle*.

(10) a. SRC But *acceptance* is something that takes time. It always takes time .

b. W/O AR Mais l'acceptation est quelque chose qui prend le temps. *Il* prend toujours le temps.

c. W/ AR Mais *l'acceptation* est quelque chose qui prend le temps. *Elle* prend toujours le temps.

The manual evaluation also revealed that refining our rules to translate cases such as (6) and (7) as *ce* instead of *il* would be a good start for tackling the under-generation problem.

Pronoun translation diagnosis	#	%
faulty parsing	12	6
non-generation of ça/cela	23	11
non-generation of ce	30	14
non-generation of translation	9	4
wrong antecedent	42	20
wrong identification of pleonastic	1	0
accurate translation	95	45
Total	212	100

Table 5.4: Results of the evaluation using oracle annotations of Its-2 translations.

Error analysis of the translations using gold standard annotations

The manual evaluations of the systems output produced during the official evaluation process of the shared task were released some time after the competition was finished (Hardmeier, Nakov, et al. 2015). The fill-in-the-gap method of evaluation ensures that errors due to agreement mismatch with a reference antecedent do not occur. It also ensures that systems are evaluated in the context of the translations they produced. In this section, we use these manual annotations to assess particular difficulties that our system encountered. We took advantage of this to re-evaluate the errors of our system and gain a deeper understanding of the causes for errors. Table 5.4 presents our analysis of the type of errors encountered.

Only 20% of the evaluated translations have an agreement problem due to an incorrect antecedent. The other major problem with the system concerns generation errors which amount to 29% of the translations. The non-generation of the French pronouns *ça*, *cela* is particularly interesting. This pronoun is mostly used for anphoric event reference, i.e., an event referenced by the previous clause content. Most of the time, it is difficult to pinpoint a discrete antecedent for this type of pronoun. This type of reference is not accounted for within the principles of the Binding Theory and therefore no generation rules were foreseen in the system. Example (11) includes some context to illustrate this problem. The *it* pronoun in (11a) refers back to the very first sentence, specifically to the series of events collected by *what we did*. Our system, on the other hand, produces *il* as a default translation, resulting in atypical non-fluent French.

(11) a. SRC So what we did was we took this bone marrow, grew up the stem cells in the lab, and then injected them back into the vein. I'm making this sound

really simple. It took five years off a lot of people, okay? And *it* put gray hair on me and caused all kinds of issues.

b. W/ AR Et *il* a mis des cheveux gris sur moi et a provoqué toutes sortes de questions.

c. FILLED PRONOUN: ça/cela

5.2.2 Resolution of null subjects

In this section, we test the AR component of the Its-2 system on the resolution of Spanish null subjects and their translation into French. Null subjects refer to omitted subject pronouns licensed in some languages. They rely on their distinctive verbal morphology to distinguish grammatical persons. An example is presented in (12).

(12) Han prometido un mejor servicio.
NULL have.3.pl promised a better service
'They have promised a better service.'

Many grammars assume the existence of a theoretical *pro* filling the empty position, giving the languages with this characteristic the name of *pro-drop* languages. This *pro* pronoun has all the properties of regular pronouns. This trait enables systems with some sort of grammar behind to handle null subjects with ease. We think that SMT systems, in contrast, could have a hard time generating full pronouns when translating from a pro-drop into a non-pro-drop language.

Evaluation of AR compononent on Spanish null subjects

We manually evaluated the translation of null subjects from Spanish to French. The AR component is the same used above for the translation from English to French. Since the AR is build within a rule-based parser, null subjects have a *pro* representation equivalent to that of overt pronouns.

We took the data for the evaluation from the AncoraES-Co corpus (Recasens and Martí 2010). This corpus is composed of newspaper articles and is annotated with coreference links, providing us with concrete antecedents for the pronouns of interest. We selected 18 articles, amounting to 250 sentences. We kept the structure of each article without changing the sentence order. 78 null pronouns and 14 non-null personal pro-

nouns were found in total, all referential pronouns (Loáiciga 2013).

Table 5.5 shows the translation results obtained with and without the AR component. We considered a translation correct when the pronoun is generated with the corresponding grammatical features of its antecedent; otherwise, the translation is considered incorrect. We obtained encouraging results with a significant improvement for the translations of null pronouns, which rose from 9 correct to 40 translations.[3]

Personal Pronoun	**Without AR**		**With AR**	
	Correct	**Incorrect**	**Correct**	**Incorrect**
Null	9	69	40	38
Non-Null	9	5	9	5
Total	18	74	49	43

Table 5.5: Results of the manual evaluation of the resolution of Spanish null subjects.

As a drawback, however, we noted that the translation quality of relative pronouns decreased. At transfer time, the identified *pro* pronouns get replaced with full pronouns. In presence of a subject pronoun, the system generates the relative pronoun with accusative case (*que*) instead of the nominative case pronoun (*qui*).

5.3 Cross-lingual pronoun prediction

5.3.1 Introduction

In the previous sections of this chapter, we have used the Its-2 machine translation system (Wehrli, Nerima, and Scherrer 2009) with a classic anaphora resolution procedure. While this approach proved effective to generate the gender inflected third person referential pronouns *il*, *ils*, *elle*, *elles*, these pronouns correspond only to a subset of the translation possibilities of English *it* and *they*. This solution is limited and confirmed that the problem of pronoun translation goes beyond the anaphora resolution problem. However, to write rules which are sufficiently general to account for all pronoun translation possibilities is perhaps impossible. We do not have enough understanding yet of the linguistic changes from source into target language behind all the translation possibilities listed in the last chapter in Table 4.4 or even in Table 4.5.

[3] χ^2 (1, N = 156) = 28.59, p <.05

Cross-lingual pronoun prediction is a classification approach to estimate a pronoun's translation directly, without generating a full translation of the segment containing the pronoun. Unlike full machine translation, the pronoun prediction task has access to both source and target language data (excepting the target pronoun) during training and testing time. This approach has the advantage of a simple and efficient modeling. On the one hand, the target translation possibilities can be defined as a fixed number of classes to predict. On the other hand, multiple sources of information can be investigated in the form of features. Cross-lingual pronoun prediction is a valuable task in its own right, although recently it has gained strength as an alternative to a full machine translation pipeline to study the translation of pronouns due to its modeling and evaluation advantages (Hardmeier, Tiedemann, and Nivre 2013).

The work presented in this section is based on our submissions to the shared tasks on cross-lingual pronoun prediction held in 2015 (Hardmeier, Nakov, et al. 2015) and in 2016 (Guillou, Hardmeier, Nakov, et al. 2016). We report the results of four experiments on cross-lingual pronoun prediction using the materials provided by the shared tasks organizers that are described below. Our experiments have a twofold goal. First, we seek to have a better understanding of the role of different kinds of linguistic information in determining a pronoun's translation. In particular, we will assess three different clusters of features: syntactical, morphological and contextual. We think that previous research on cross-lingual pronoun prediction is based on features certainly relevant for the task, but there is no systematic link between the types of features and the interaction they may have with each individual class to predict. Second, we assess the performance of the classifiers for the task without using explicit anaphora or coreference resolution knowledge. Our motivation for the second aspect is that anaphora and coreference resolution systems introduce many errors due to the heavy pre-processing they rely on, and to their own performance in matching pronoun-antecedent pairs. Besides, most of them exist only for English. Last, most of these systems provide coreference links only for a subset of the pronouns we are interested in, as proven by our own experiments in this line.

Both cross-lingual pronoun prediction shared-tasks were defined as fill-in-the-gap tasks: given an input text and a translation with placeholders, replace the placeholders with pronouns. All experiments reported here refer to the English-French translation pair. There are nine defined prediction classes, *ce, cela, elle, elles, il, ils, on, ça* and OTHER, the last one being a catchall for translations as lexical NPs, as reformulations and as

paraphrases or with no pronoun at all. The systems were evaluated as in a standard classification task. As training data, the participants were given the English source, the translation with gaps and bidirectional word alignments. An example of the provided data is given in Figure 5.1.

SOURCE	Even though *they* were labeled whale meat , *they* were dolphin meat .
TRANSLATION	Même si REPLACE_2 avaient été étiquettés viande de baleine , REPLACE_8 était de la viande de dauphin .
ALIGNMENT	0-0 1-1 2-2 3-3 3-4 4-5 5-8 6-6 6-7 7-9 8-10 9-11 10-16 11-13 11-14 12-17
CLASSES	*ils, ce/c'*

Figure 5.1: Example of training data for the 2015 shared task on cross-lingual pronoun prediction.

The baseline provided consists in the predictions of a 5-gram language model (LM) trained on the complete training data (TED talks from the WIT^3 project (Cettolo, Girardi, and Federico 2012), Europarl version 7 (Koehn 2005), News Commentary version 9 and news data from WMT 2007–2013 (Bojar, Buck, Federmann, et al. 2014)). The LM fills the gaps with each of the eight target pronouns classes on the one hand, and with the most frequent words in the corpus (22 tokens) including NONE (not inserting anything) on the other hand to account for the OTHER class. The filler which produces the sequence with the higher probability is picked as the predicted class.

The task held in 2016 was very similar to the 2015 one with the exception of two aspects. Firstly, instead of a single English-French task, three more subtasks were opened: French-English, English-German and German-English. Secondly, the baseline was not built using fully inflected target language but lemmatized and POS-tagged text. This had the purpose of approximating the machine translation scenario even more, since morphological information that would reveal certain agreements is removed. An example of the provided data for the 2016 task is presented in Figure 5.2.

5.3.2 Data and tools

As a preprocessing step, both sides of the parallel data are parsed using the rule-based Fips parser (Wehrli 2007). This parser produces an information-rich phrase-structure representation with predicate-argument labels. Besides, it can also be used as a tagger,

SOURCE	Even though *they* were labeled whale meat , *they* were dolphin meat .
TRANSLATION	même|ADV si|KON REPLACE_2 avoir|VER être|VER étiqueter|VER viande|NOM de|PRP baleine|NOM ,|PUN REPLACE_8 être|VER de|PRP la/le|PRON viande|NOM de|PRP dauphin|NOM .|.
ALIGNMENT	0-0 1-1 2-2 3-3 3-4 4-5 5-8 6-6 6-7 7-9 8-10 9-11 10-16 11-13 11-14 12-17
CLASSES	*ils, ce/c'*

Figure 5.2: Example of training data for the 2016 shared task on cross-lingual pronoun prediction.

since it generates a POS-tags (containing disambiguated morphological information) and a grammatical functions for each word of a given sentence. We relied on this tagger output for extracting most of our features. An example of the output is given in Figure 5.3.

And	CONJ-COO	and	
it	PRO-PER-3-SIN	it	SU
's	VERB-IND-PRE-3-SIN	be	
a	DET-SIN-NEU	a	FO
very	ADV-INT	very	
easy	ADJ	easy	
question	NOUN-SIN-NEU	question	
.	PUNC-POINT		

Figure 5.3: Example of the tagger output of the Fips parser for the sentence *"And it's a very easy question"*. The first column contains the words in the sentence, the second the POS-tags and morphological analysis, the third consists of the lemmas and the fourth of the predicate-argument labels. SU stands for subject, FO stands for predicative complement.

For the French side, a unique placeholder is inserted in the place of each REPLACE_XX item. This ensures coherent syntactic analysis by the parser, since projections are based on the lexical properties of the heads. The placeholder was inserted in the lexicon as a token with all possible morphological features: both masculine and feminine gender, singular and plural number and the three possible persons. Due to its rule-based nature, the parser unifies only the compatible feature values on each sentence. Consequently, the placeholder allowed us to retrieve some information from the unification process with the verb.

The final training data used consists of 28,422 examples composed from a subset of the shared task data. It includes 747 instances from the TED talks, 14,561 from News Commentary and 13,114 from Europarl. All systems are built using the Stanford Maximum Entropy package, MAXENT. The distribution of the classes in the training data is presented in Table 5.6 Note that the OTHER class has overwhelmingly high frequency with respect to the other classes.

Class	Examples
ce	2,326
cela	694
elle	2,786
elles	1,101
il	6,358
ils	3,623
on	516
ça	83
OTHER	10,935
Total	28,422

Table 5.6: Distribution of the classes in the training data for experiments 1 to 4.

5.3.3 Features

We use three types of features roughly following the categorization of Friedrich and Palmer (2014). Most of them rely on the predicate-argument structure of the English side and morphological analysis of the French side. The rationale for this choice is to simulate an MT scenario (where target sentences are not available) in which one could parse the source language to find the argument of interest and may use a dictionary for getting the target-language correspondent morphology.

For each training example, we extracted syntactic, morphological and contextual information as the features 1 to 18 below. The possible values of all features are listed in Table 5.7. The distance between a pronoun and its antecedent is implicitly handled by a language model within a limited window when computing n-gram probabilities. In an attempt to model the notion of distance between the pronoun and each of the arguments in the sentence, we experimented with the position of each argument as a feature. This did not change anything to the model, therefore we dropped it early on.

1. Current sentence subject

2. Current sentence object
3. Current sentence predicative object
4. Current sentence sentential object
5. Previous sentence subject
6. Previous sentence object
7. Previous sentence predicative object
8. Previous sentence sentential object
9. Gender and number of all adjectives
10. Previous word POS-tag
11. Following word POS-tag
12. Voice of following verb
13. Person and number of following verb
14. Previous lemma
15. Following lemma
16. Previous word token
17. Following word token
18. Second following word token

The **syntactic features** 1 to 8 refer to the arguments present in the English sentence (fourth column in Figure 5.3). Once an argument is identified in the English sentence, the gender and number of the word-aligned French token (most often the head) is retrieved. In the case of the sentential objects, only the values YES or NO are assigned. Sentential objects are sentences acting as complements of the verb and very often with a conjunction or preposition as their head; therefore, we did not look for gender and number.

Morphological features 9 to 13 concern the POS and morphological tags of the words in the immediate context of each pronoun to predict (second column in Figure 5.3). To obtain the value for feature 9, all adjectives in the previous and the current sentence are identified and the gender and number of their French word-aligned token is searched. Then French gender and number information is aggregated and the most frequent one is

selected.

Finally, the **contextual features** 14 to 18 refer to the preceding or following tokens of each French pronoun to predict. For these, sentences are concatenated and their boundaries are ignored. For instance, if the previous word happened to be the full stop of the previous sentence, a full stop is then taken as the value for previous word token.

Features	**Values**
1,2,3,5,6,7,9	{SIN-FEM, SIN-MAS, PLU-FEM, PLU-MAS, INN-FEM, INN-MAS}
4,8	{YES, NO}
10,11	{NOUN, VERB, ADV, PRO, CONJ, PUNC, DET, ADJ, PREP}
12	{ACTIVE, PASSIVE}
13	{1-SIN, 1-PLU, 2-SIN, 2-PLU, 3-SIN, 3-PLU}
14,15	e.g. {*le, avoir, venir, être, rester, ...*}
16,17,18	e.g. {*la, ont, viennent, sont, restent, ...*}

Table 5.7: Possible values for each of the features. INN stands for *unknown number*.

5.3.4 Experiment 1: Cross-lingual pronoun prediction

In this experiment, all the features previously introduced are used. However, features 1 and 5 refer to subjects, which are likely to be pronouns aligned with REPLACE_XX items on the French side. In order to simulate the use of an unmodified parser, we dropped the morphological features obtained by unification for the REPLACE_XX items and inserted the special feature value PRON instead. Table 5.8 contains the results of the experiment.

5.3.5 Experiment 2: Cross-lingual pronoun prediction with unification values

For this second experiment, we use the unified values for REPLACE_XX subjects (features 1 and 5). Additionally, the vast OTHER class was split in two classes in order to reduce the imbalance: i) translations by a pronoun not considered among the classes or by a lexical NP (i.e., genuine OTHER), and ii) translations without anything in French (i.e., NONE). The labels for the latter were taken from the annotation provided with the training data. After classification, the two sub-classes were merged again. The obtained results are presented in Table 5.8.

A comparison of the results of both experiments is depicted in Figure 5.4. Differences between the results reported in Loáiciga (2015) and those here are due to the correction of a bug after the system description paper was published and while working on Experiments 3 and 4 below. The results reported here are slightly higher but they do not change the overall ranking obtained in the shared task (9th place among 14 submitted systems).

Class	**Experiment 1**				**Experiment 2**			
	S+M	**S+C**	**M+C**	**S+M+C**	**S+M**	**S+C**	**M+C**	**S+M+C**
ce	19.38	**73.49**	75.00	71.30	23.39	***75.22**	74.70	73.87
cela	0	**13.56**	6.90	12.12	0	***16.39**	9.68	12.70
elle	15.65	37.84	40.44	**42.04**	28.57	36.99	38.34	**42.24**
elles	28.17	**35.14**	32.91	34.15	28.57	***37.33**	23.28	34.57
il	30.30	44.88	42.58	**46.35**	31.35	***49.28**	41.18	48.61
ils	68.57	78.26	79.75	**80.37**	70.51	77.22	77.44	***80.62**
on	0	36.92	36.67	***38.71**	7.27	30.59	33.90	**35.29**
ça	0	***14.41**	14.29	14.29	0	**11.11**	5.61	10.91
OTHER	79.80	83.42	***87.27**	86.43	79.63	83.00	**86.06**	85.57
	Accuracy: 721/1105 (65.25%)				**Accuracy:** 724/1105 (65.52%)			

Table 5.8: Comparison of F1 scores (%) obtained in Experiments 1 and 2 with different groups of features. **S** stands for syntactic, **M** stands for morphological and **C** stands for contextual features. The best results per class within experiments are presented in **bold**, while the best results per class across experiments are marked with stars (*). Accuracy refers to the **S+M+C** classifiers.

5.3.6 Discussion

From both Table 5.8 and Figure 5.4, it can be noted that most of the results of the system 2 are better than those of system 1. It can additionally be noted that the absence of morphological features (column **S+C** vs **S+M+C** in Table 5.8) seems to have a rather small impact on the final results. The syntactic features, in contrast, do seem to be important. They are motivated in the salience hierarchies established within linguistic theories of salience and AR. In these theories, a syntactically salient argument such as the subject is more likely to be the antecedent of a pronoun. Our results show that this particular set of features contributes much of the knowledge to the model. This finding seems to be consistent with the results reported by Stymne (2016) who found that dependency links on the pronouns and their heads were among the most useful features for English-French and English-German cross-lingual pronoun prediction. This finding, however,

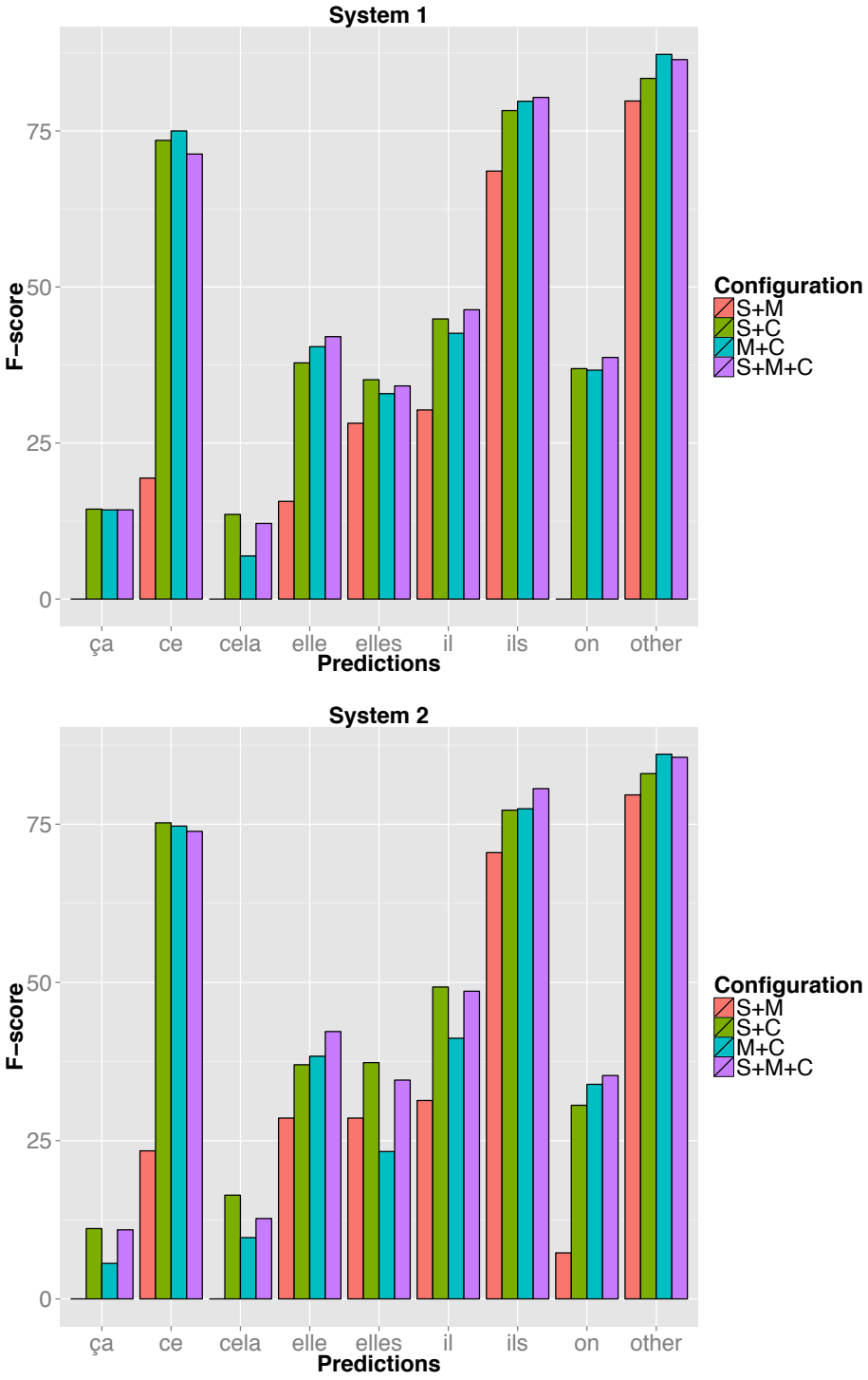


Figure 5.4: Comparison of F1 scores (%) obtained in the test set with different groups of features in Experiment 1 and Experiment 2.

differs from Kehler et al. (2004)'s results which showed that predicate-argument statistics (subject-verb, verb-object, possessive-noun) as observed in a large corpus are of very little benefit to the task of pronoun resolution. In their work, the authors exploited the relative frequency of the predicate-argument preferences selections. They give the example of the sentence in (13) where *industries* is observed more frequently in corpora as the head of the object noun phrase of *force* than *edge*, making it a more likely antecedent.

(13) He worries that Glendening's initiative could push his industry over the edge, forcing *it* to shift operations elsewhere.

Turning now to the morphological features, they do not seem to influence the classifier decisions, and in some cases they only add noise, as indicated by the drop in performance in columns **S+M** and **M+C**. Although they do not do much on their own, they have some weight in combination with the other two groups of features (columns **S+M+C**). In particular, they add some gain to the pronouns *ils, elle* and *on*, the last two being rather difficult, as evidenced by their lower scores in comparison to *ils*. This result is also sensible for the *ils* and *elle*, since they are inflected for gender and number in French. As for the pronoun *on*, the gain reflects perhaps our observation (Section 4.3.1) that many *on* occurrences are translations of English passive constructions of verbs with two objects such as *to give* in (14). Pronoun *ça* is the class that benefits least from the morphological features.

(14) a. EN. And gamers are willing to work hard all the time, if *they're given* the right work.
b. FR. Les joueurs sont prêts à travailler dur, tout le temps, si l'*on* leur confie la bonne mission.

Both systems additionally show that contextual features are highly important. When they are removed from the model (columns **S+M**), a substantial drop in performance is observed for all the classes, in particular for the first system, which do not have the subject information from unification (Figure 5.4). Contextual features are particularly determinant for the *ça* and *cela* pronouns. We expected that these pronouns were mostly governed by sentential objects instead, either from the current or the previous sentence. The importance of the contextual features partially explains the strength of the baseline

System	Feature number
System 1	13,18,10,17,15,16,14,12,11, 8,2,3,5,9,6,7,1,4
System 2	13,18,17,10,15,16,14,11,12 1,4,8,2,9,5,3,6,7

Table 5.9: Features of the model ordered from the most to the least informative according to accuracy.

system provided in the shared task (which ranked 1st among the 14 submissions).

Looking at the features individually (Table 5.9), it can be noted that for both systems the morphology information of the following verb (feature 13) is the most important parameter, which is not very surprising since the task deals mostly with subject pronouns. The other top-ranking features are the following word POS-tag (18), the following lemma (17) and the previous predicative object (10). Predications are indeed very informative since they agree with the subject in French. This feature may be less useful in a target language like English or German, where this is not the case.

The hierarchy in Table 5.9 reveals further understanding about the context features as well. Features concerning lemmas (15 and 14) have almost as much weight as features concerning raw tokens (16, 17, 18), especially the following lemma. Their influence depends on the pronoun to predict: while raw tokens are good predictors for pronouns *ce, ça* and *on*, lemmas are good predictors for pronouns *il, elle, ils* and *elles*. Contrary to intuition, this pattern suggests that the immediate inflected context offers little clue for the prediction of *il, elle, ils* and *elles*, which only differ in their gender and number inflections. These pronouns in particular seem to need explicit knowledge of their antecedents for their adequate translation. We tried to approximate the antecedent knowledge which may be provided by an AR system with the syntactic features, but our approach work less well than expected. The system submitted by Wetzel, Lopez, and Webber (2015), on the other hand, uses very similar features to our own and the same classifier, but it also includes antecedent information from an AR system. Their submission obtained better results for pronouns *il*, *ils* and, in particular, *elles*.[4]

As mentioned and depicted in Figure 5.5, results from our System 2 are better[5] than those of System 1 for all the classes. This result evidences a bias towards the OTHER class in

[4]They obtained F-scores of 54.39 for *il*, 41.73 for *elle*, 83.43 for *ils* and 47.89 for *elles*.

[5]τ = -12.1579, df = 1104, p-value < 2.2e-16

System 1, harming the less frequent classes. Our two-way distinction is straightforward using the provided data, but we suspect that a finer distinction could further improve results. One could for instance distinguish between subject pronouns and object pronouns.

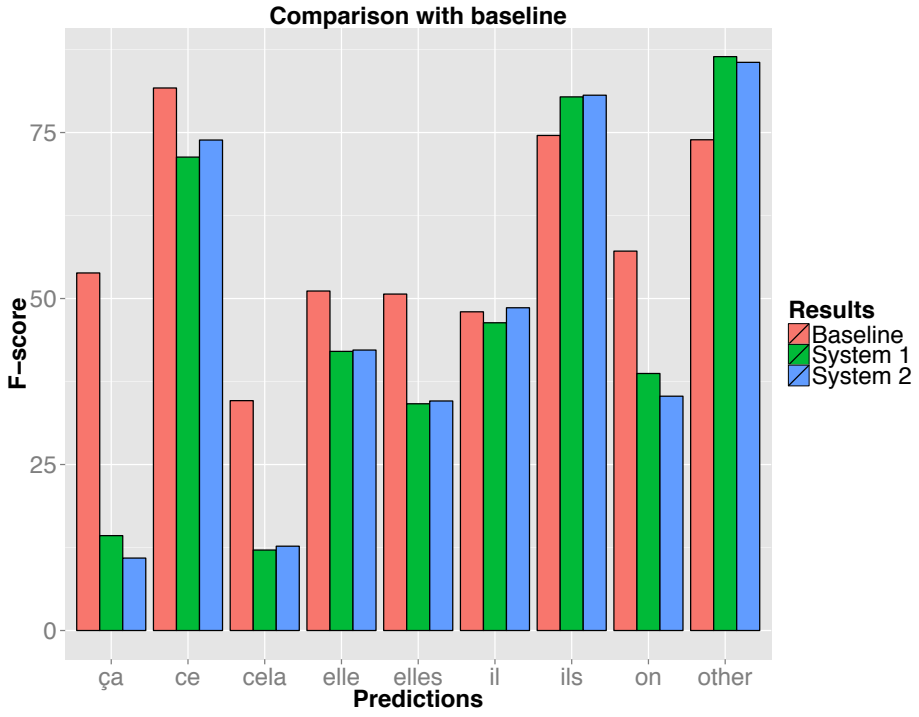


Figure 5.5: Comparison of fine-grained F-scores of the System 1 and System 2 with the shared task baseline.

Figure 5.5 also shows that our systems underperform in comparison with the baseline. It is worth pointing out, however, that none of the submitted systems outperformed the LM baseline in the 2015 shared task. The only source of information available to language models is context. In addition, since contextual features proved to be the best predictors for our two previous experiments, the following two experiments are focused on understanding the strengths of the n-gram knowledge contained in the baseline LM. We explore two different ways in which such knowledge can be exploited to improve the performance of our classifier. In the first experiment, we include the predictions of the LM as part of the features fed to the MAXENT classifier. In the second, we do a log-linear system combination of the LM and the MAXENT classifier.

5.3.7 Experiment 3: LM predictions as features

This experiment and the next one make use of the same classifier presented in Experiment 1 and trained on the combination of the syntactic, morphological and contextual features (column **S+M+C** in Table 5.8). Results are therefore directly comparable. In this experiment, the LM provided as the shared task baseline is queried and the predicted pronoun token is added to the feature vector of each instance to classify. The LM is queried with each of the eight target pronouns and with a list of the most frequent words in the corpus (22 tokens) to simulate the OTHER class. Results are presented below in Table (5.10).

Class	**Precision**	**Recall**	**F1**	**Δ F1**	**Exp. 1**	**LM**
ce	90.00	83.15	86.44	+15.14	71.30	81.71
cela	28.57	37.04	32.26	+20.14	12.12	34.62
elle	45.68	44.58	45.12	+3.08	42.04	51.13
elles	75.76	49.02	59.52	+25.37	34.15	50.67
il	48.12	74.04	58.33	+11.98	46.35	48.00
ils	84.76	86.88	85.80	+5.43	80.37	74.56
on	51.61	43.24	47.06	+8.35	38.71	57.14
ça	76.47	25.49	38.24	+23.95	14.29	53.85
OTHER	83.12	92.4	86.43	+1.1	73.91	87.53
Accuracy: 813/1105 (73.57%)				(+8.32%)	721/1105	733/1105

Table 5.10: Results obtained (%) after integrating the baseline predictions as additional features to the classifier presented in Experiment 1. The **Δ F1** column presents the gain in performance with respect to the results of Experiment 1. The column **Exp. 1** duplicates the results of Experiment 1 to facilitate comparison. The column **LM** presents the results using LM predictions only.

The performance of this new system improves in all the classes. There is an average gain of 12.72% F1 score in performance from the combination of the classifier from Experiment 1 with the LM predictions. In comparison with the performance obtained by the language model alone, the new system obtains improvements for the *ce*, *elles*, *il* and *ils* classes. Although, as we mentioned before, the immediate context (one token to the left and two tokens to the right) offered little clue for the prediction of *il, elle, ils* and *elles*, this experiment suggests that the original system would have benefitted from a bigger context (five tokens) for the correct prediction of these same classes.

5.3.8 Experiment 4: Log-linear combination of LM and MAXENT

An alternative method to exploit the baseline language model predictions is to combine them with the classifier by interpolation of the probability distributions they produce (5.1). The outcome is a new classification decision formed by the strengths of each of the systems. For an uniform combination, one could give each of the systems 50% of the weight in the decision. However, we opted for an informed system combination and tried different values of λ.

$$p(i) = \lambda \cdot \log p_{\text{MAXENT}}(i) + (1 - \lambda) \cdot \log p_{\text{LM}}(i) \tag{5.1}$$

As mentioned earlier, two groups of fillers are used to query the LM for a particular prediction. One group is composed of the target pronouns or classes and the other one is composed of the most frequent words in the corpus (22 tokens) to account for the OTHER class. However, in this experiment we were interested in obtaining the probability distribution of the classes of interest. Since *c'* and *ce* are treated as different words by the LM, but they are mutually exclusive, we sum their probabilities. Also, to obtain a probability for the OTHER class, we took the result of 1 minus the sum of all other pronoun probabilities.

Given that all the training data and the development data were included in the LM training by the shared task organizers (Hardmeier, Nakov, et al. 2015), and the development set is also part of our MAXENT classifier training data, the parameters for the system combination were optimized using the test set data directly. These results give a direct comparison with the results of the systems built in the previous experiments, however, ideally we would have used a different development set. To test the predictive power of the system combination on new unseen data, we also report the results computed on the 2016 shared task test set.[6]

The optimization process of λ was done using accuracy. F-score was avoided since it penalizes errors twice, once in the computation of precision and again in the computation of recall. Since only two parameters are involved, we did exhaustive search with sampling every 0.1 (Table 5.11). Figure 5.6 depicts the process at the different values of λ.

[6]The 2016 shared task used lemmas in the target-side data. We obtained the inflected data necessary to extract the features for our classifier and to query the LM from the organizers.

Configuration	Accuracy
$\lambda = 1.0$ (MAXENT)	65.25
$\lambda = 0.9$	66.15
$\lambda = 0.8$	68.05
$\lambda = 0.7$	68.69
$\lambda = 0.6$	70.41
$\lambda = 0.5$	69.95
$\lambda = 0.4$	70.32
$\lambda = 0.3$	70.41
$\lambda = 0.2$	70.41
$\lambda = 0.1$	69.68
$\lambda = 0.0$ (LM)	66.33

Table 5.11: Results of the optimization process with different values of λ.

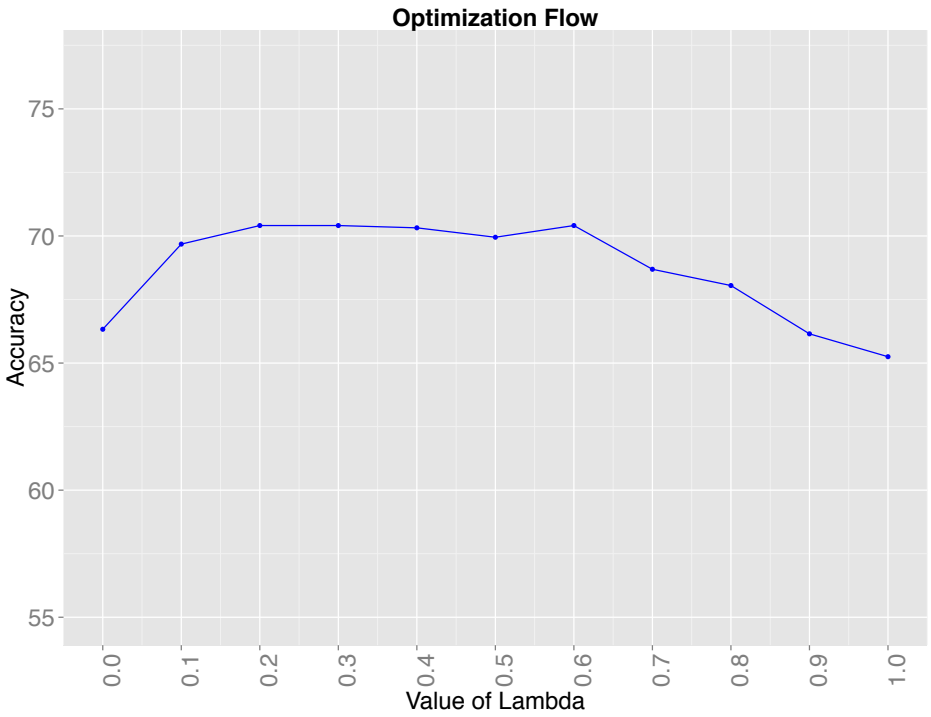


Figure 5.6: Optimization process at different values of λ.

5.3.9 Further discussion

The comparison of the results from the last two experiments (Table 5.10, Table 5.12 and Table 5.13) reveals that including the language model predictions in the form of additional features to the training of the classifier produces the biggest increments in

Class	Precision	Recall	F1	Δ F1	Exp. 1	LM
ce	87.27	78.26	82.52	+11.22	71.30	81.71
cela	21.88	25.93	23.73	+11.61	12.12	34.62
elle	60.29	49.40	54.30	+12.26	42.04	51.13
elles	72.73	31.37	43.84	+9.69	34.15	50.67
il	40.43	73.08	52.05	+5.67	46.35	48.00
ils	80.72	83.75	82.21	+1.84	80.37	74.56
on	59.09	35.14	44.07	+5.36	38.71	57.14
ça	91.67	10.78	19.30	+5.01	14.29	53.85
OTHER	78.14	94.12	85.39	-1.04	73.91	87.53
Accuracy: 778/1105 (70.41%)				(+5.16%)	721/1105	733/1105

Table 5.12: Development results of the system combination of the classifier presented in Experiment 1 and the language model given as baseline with $\lambda = 0.6$.

Class	Precision	Recall	F1	Δ F1	MaxEnt	LM
ce	95.45	30.88	46.67	+12.94	33.73	82.17
cela/ça	71.43	16.13	26.32	+1.93	24.39	11.76
elle	68.00	73.91	70.83	+30.83	40.00	60.00
elles	81.82	36.00	50.00	+5.00	45.00	25.81
il	50.00	85.25	63.03	+7.35	55.68	61.76
ils	80.00	84.51	82.19	+6.85	75.34	69.23
on	20.00	22.22	21.05	-1.17	22.22	52.17
OTHER	68.91	96.47	80.39	-0.86	81.25	74.64
Accuracy: 248/373 (66.49%)				(+6.97%)	222/373	242/373

Table 5.13: Results of the system combination of the classifier presented in Experiment 1 (column **MaxEnt**) and the language model given as baseline (column **LM**) using interpolation of their probability distributions with $\lambda = 0.6$. Results computed on the 2016 shared task test-set. The **Δ F1** column shows the difference with respect to the results using only the classifier. In this test set, *ça* and *cela* are merged into a single class.

performance. This outcome is clearly manifested in the green columns in Figure 5.7.

The system built in Experiment 4, on the other hand, adds roughly the same gain to all the classes. The gains are somewhat smaller than those of Experiment 3. However, there is a comparatively large gain for the low frequency classes; the *elle* class in particular shows an increment of 12.26 F1 vs 3.08 F1 in Experiment 3. The same is observed in the results computed on the 2016 test set, where the *elle* class gains 30.83 F1 points. This outcome suggests that this particular method directly alleviates the weaknesses of the classifier in some of the classes with low frequency.

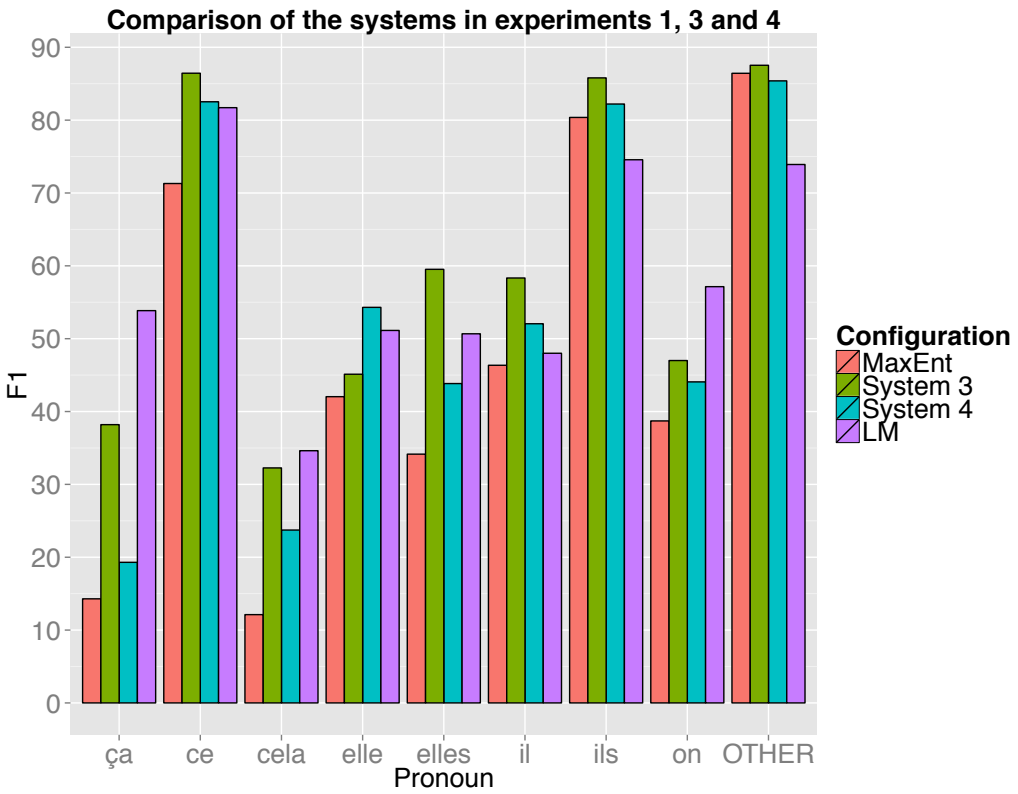


Figure 5.7: Comparison of fine-grained F-scores of the combined systems and the task baseline.

The results of these experiments underline once again the importance of the contextual information, in this case provided by the LM. However, we have also shown that syntactic features were relevant for the task of pronoun prediction. In this sense, the results from the last two experiments indicate that a combination of both sources of information produces the best outcomes. Both system combinations consistently outperform each of the systems for the classes *il, ils* and *elles*, which are strongly dependent on a nominal antecedent. Taken independently, system 3 outperformed the baseline for the classes *ce, elle, il, ils* and system 4 for the classes *cela/ça, elle, elles, il, ils* and OTHER (Table 5.13).

Other participants to the shared task included features from an external anaphora or coreference resolution system but did not beat the baseline either. In this context, our last two experiments suggest that a combination of syntactic and (enough) contextual information offers an alternative to these systems for the task of pronoun prediction.

Looking at Figure 5.7, it is clear that there are two levels of difficulty. The classes *cela,*

elle, elles, il, on and *ça* are more difficult to predict by all the tested configurations than the classes *ce, ils* and OTHER. These last three classes can be predicted based on the surrounding context mostly.

To sum up, the combined systems are the best option overall. Neither the classifier nor the LM perform as well as the the combination of the two. In addition, it is to note that the *ça* and *cela* classes present the worst performance figures throughout all of our experiments. These pronouns are referential in French, but they do not point to a nominal antecedent. In the following chapter, we address this problem from a different perspective.

5.4 Conclusion

In this chapter, we have investigated two different approaches for pronoun translation from English into French: rule-based translation with classic anaphora resolution and cross-lingual pronoun prediction without anaphora resolution.

The first exploits the syntactic knowledge intrinsic to the parser the translation system is built on to find antecedents for the pronouns. This resulted in accurate translations for the cases in which a pronoun has a nominal antecedent and the system is able to find it. But it also provided evidence to the fact that not all pronouns are translated by a pronoun of the same category, partly because not all of them have a nominal antecedent. Given what we know about the distribution of the translation of pronouns, it is difficult to create enough rules that generalize all the translation possibilities in all possible contexts.

Our cross-lingual pronoun prediction experiments, on the other hand, have allowed us to model different types of information and to test their predictive power over a limited number of classes. Using this approach, we have also provided evidence in favour of including syntactic knowledge for the task. We have found that syntactic information combined with enough contextual information could represent an alternative to the information provided by external anaphora and coreference resolution systems.

Chapter 6

Disambiguation of *it*: prediction of event pronouns

6.1 Introduction

Our experiments in the preceding chapter have focused on the translation of the subject pronouns *it* and *they* from English into French. Both of these pronouns have multiple translation possibilities partly based on the functions they perform in text. The disambiguation of their function is required if they are to be translated correctly into other languages (Guillou 2016). For example, the pronoun *they* is typically used as an anaphoric pronoun, but may also be used generically, as in '*They say it always rains in Scotland*'. The pronoun *it* may be used as pleonastic or for nominal and event anaphoric reference. Examples of the pronoun functions treated in this chapter are provided in Figure 6.1.

PLEONASTIC	*It* is raining.
NOMINAL ANAPHORA	I have a bicycle. *It* is red.
EVENTUAL ANAPHORA	He lost his job. *It* came as a total surprise.

Figure 6.1: Examples of different pronoun functions.

Pronouns used in *nominal anaphora* corefer with a noun phrase (i.e. the *antecedent*). *Pleonastic* pronouns, in contrast, do not refer to anything but are required to fill the subject position in many languages, including English, French and German. Pronouns used for *eventual anaphora* corefer with a verb, verb phrase, clause or even with an entire sentence. Different pronouns are required when translating an instance of *it* depending on its

function and the language we are translating into. In French, for example, the anaphoric *it* may be translated with the third-person singular pronouns *il* [masc.], *elle* [fem.], and less frequently, with the plurals *ils*, *elles*, or with an non-gendered demonstrative such as *ce* and *cela*. The French pronoun *ce* may function both as event reference and as a pleonastic pronoun, but English *it* pronouns with a pleonastic function are translated only as *il*.

The translation of pleonastic and event reference pronouns poses a particular problem for MT systems (Guillou, Hardmeier, Nakov, et al. 2016). Poor performance may be attributed to the inability of the systems to disambiguate the various possible functions of the pronoun *it*. For instance, the low scores achieved for the pronouns *ça* and *cela* in our own experiments reported in the previous chapter seem to suggest a problem with the translation of pronouns with event reference function. In Figure 6.2, we show the translation distribution of 58 *it* pronouns that we labeled manually with event reference function. The data comes from the test set provided for the 2016 shared task on cross-lingual pronoun prediction (Guillou, Hardmeier, Nakov, et al. 2016). This figure shows that while in French this pronoun function can have five translations, in German it has four, with a different distribution. In French, it mainly corresponds to either *ce* or *cela*, while in German it is most frequently translated as *es* and OTHER.

Moreover, coreference resolution systems such as the Stanford CoreNLP (H. Lee et al. 2011) often include heuristics to recognize pleonastic *it*, to avoid pointless attempts to find their antecedents. The coreference resolution task can be seen as a two-step problem: mention identification followed by antecedent identification. Identifying instances of pleonastic *it* typically takes place in the first step. The recognition of event reference *it*, however, is not currently included in these systems, although it would be advantageous to incorporate it in the second step, as suggested by T. Lee, Lutz, and Choi (2016). Systems such as NADA (Bergsma and Yarowsky 2011), on their part, specialize in distinguishing between anaphoric and non-anaphoric instances of *it*.

In this chapter, we address the problem of disambiguating the function of the English pronoun *it*. This work was first introduced in Loáiciga, Guillou, and Hardmeier (2016) and it proposes the identification of nominal anaphoric, event reference, and pleonastic instances of *it* using a single system. Our hypothesis is that this three-way distinction is beneficial for accurate pronoun translation, since, besides the language one is translating into, the translation depends on the function of the pronoun in the source language.

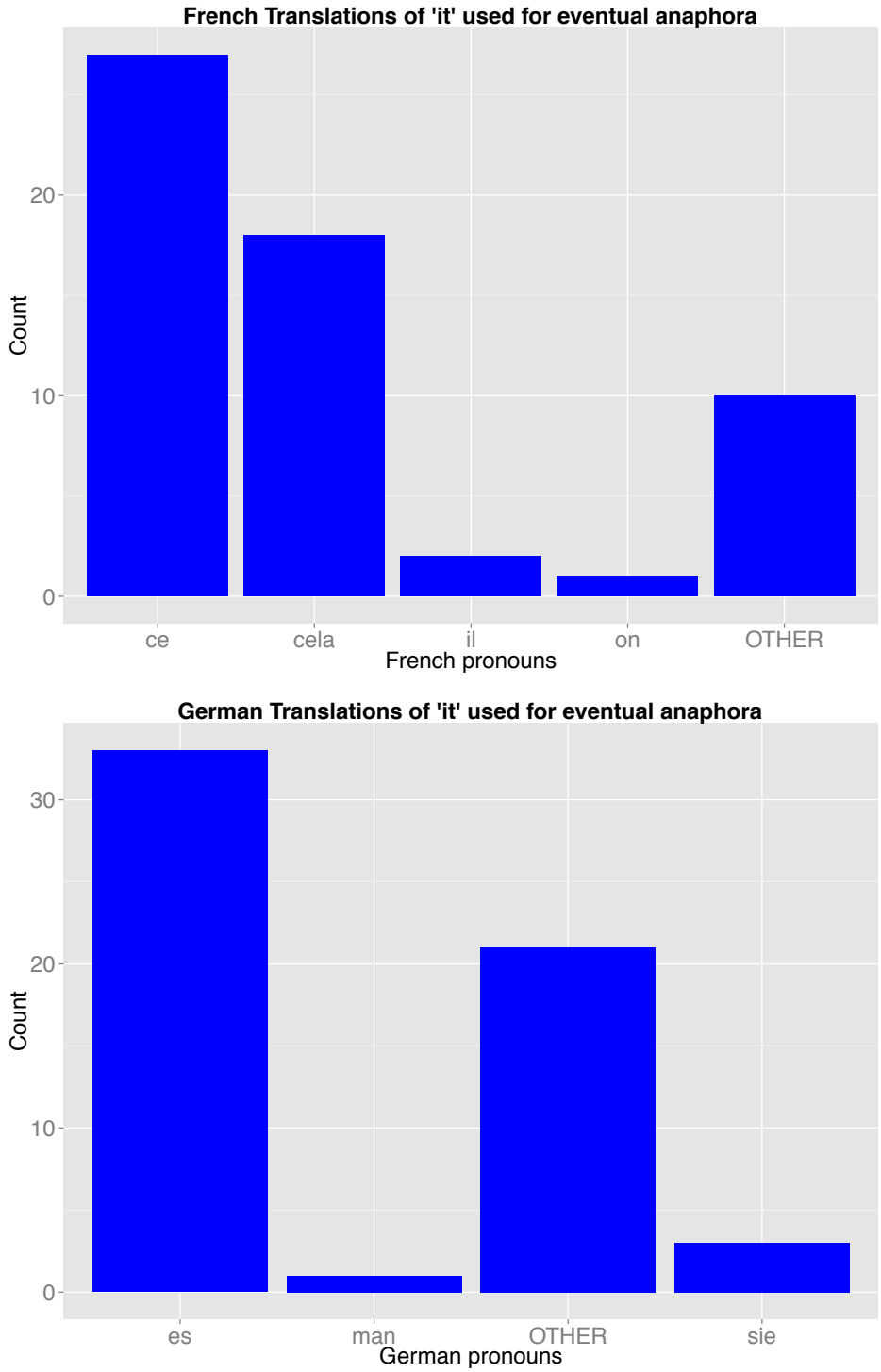


Figure 6.2: Translation distribution of French and German translation of 58 English pronouns *it* used with event reference function.

We propose classification experiments using gold-standard data and silver-standard data. The systems indicate for each instance of *it* whether the pronoun function is pleonastic, nominal anaphoric or event anaphoric reference. The main classifier was trained using data from the ParCor corpus (Guillou, Hardmeier, Smith, et al. 2014) and the *DiscoMT2015.test* dataset (Hardmeier, Tiedemann, Nakov, et al. 2016). In both corpora, pronouns are labeled according to their function, following the ParCor annotation scheme. This scheme labels pleonastic pronouns as *pleonastic*, pronouns used for nominal anaphora as *anaphoric* and pronouns used for eventual anaphora reference as *event*[1]. The classifier with the best performance is incorporated in an extended language model (LM) for the English-to-French pronoun prediction task. The extended LM is an *n*-gram language model that operates over target-language lemmas, but also has access to the source-language pronouns.

6.2 Data

The ParCor corpus and *DiscoMT2015.test* dataset were used as gold-standard data. For all instances of *it* labeled as anaphoric, pleonastic or event, the sentence-internal position of the pronoun and the sentence itself are extracted. The corpora included a number of instances of *it* labeled as being cataphoric or having extra-textual reference. These are excluded from the classifier training data. The pronouns *this* and *that*, when used as event reference pronouns, may in many cases be used interchangeably with the pronoun *it* (Guillou 2016). Consider example (1), in which the pronouns *this* and *it* may be used to express the same meaning.

(1) a. John arrived late. *This* annoyed Mary.
b. John arrived late. *It* annoyed Mary.

To increase the number of training examples, instances of event reference *this* and *that* are replaced with *it* and added to the training data. The data was divided into 1,504 instances for training, and 501 each for the development and test sets. All sentences were shuffled before the corpus was divided, promoting a balanced distribution of the classes (Table 6.1).

[1] Other categories of the corpus are not considered here.

Data Set	it-Event	it-Anaphoric	it-Pleonastic	Total
Training	504	779	221	1,504
Development	157	252	92	501
Test	169	270	62	501
Total	830	1301	375	2,506

Table 6.1: Distribution of classes in the gold-standard training data used for the *it* disambiguation experiments.

6.3 Baselines

For development and comparison purposes we built two different baselines. One is a 3-gram language model built using KenLM (Heafield 2011) and trained over a modified version of the annotated corpus in which every *it* is concatenated with its type (e.g. *it-event*). For testing, the *it* position is filled with each of the three *it-label* and the language model is queried.

Table 6.2 presents the results of this baseline using 14-fold cross-validation and a single held-out test set. The motivation for the choice of the number of folds is threefold. First, we wanted to respect document boundaries; second, we aimed for a fair proportion of the three classes in all folds; and, lastly, we tried to lessen the variance given the relatively small size of the corpus.

The second baseline (Table 6.3) is a setting in which all instances of the test set are set to the majority class *it-anaphoric*. The majority class baseline for the 14-fold cross-validation is equivalent to set all the labels in the corpus to *it-anaphoric*.

	14-fold cross-validation			Test-set		
	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.599	0.248	0.350	0.732	0.262	0.387
it-pleonastic	0.152	0.621	0.244	0.139	0.694	0.231
it-event	0.528	0.277	0.363	0.521	0.290	0.373
Accuracy	(785/2506) 0.313			(163/501) 0.325		

Table 6.2: N-gram baseline for the classification of the three types of *it*.

The 3-gram baseline appears to be biased towards the pleonastic class, as suggested by its high precision and very low recall for the event and anaphoric classes and the opposite situation for the pleonastic class. In addition, the first baseline shows that the disambiguation of the three types of *it* is difficult in an *n*-gram context such as that

	14-fold cross-validation			Test-set		
	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.519	1	0.683	0.539	1	0.700
Accuracy	(1301/2506) 0.519			(270/501) 0.539		

Table 6.3: Majority class baseline for the classification of the three types of *it*.

provided by a language model. The identification of pleonastic realisations, in particular, is very hard in this context.

6.4 Design and features

We first experimented with a Maximum Entropy classifier, MAXENT, built with the Stanford Maximum Entropy package. To extract the features fed to the classifiers, we parsed the corpus with the joint part-of-speech tagger and dependency parser of Bohnet et al. (2013) from the Mate toolkit. In addition, the corpus was lemmatized using the TreeTagger lemmatiser (Schmid 1994). Although other tools (described below) were used, we relied on the output of these two parsers to extract most of our features.

For each training example, we extract the following information:

1. Previous three tokens. This includes words and punctuation. It also includes the tokens in the previous sentence when the *it* occupies the first position of the current sentence.
2. Next two tokens
3. Lemmas of the next two tokens
4. Head word. Most of the time the head word is a verb.
5. Whether the head word takes a *that* complement (verbs only)
6. Tense of head word (verbs only). This is computed using the method described in Chapter 7, section 7.2.
7. Lemma of the head word
8. Presence of *that* complement in previous sentence. A binary feature which follows Navarretta (2004)'s conclusion (for Danish) that a particular demonstrative pronoun (*dette*) is often used to refer to the last mentioned situation in the previous

sentence, often expressed in a subordinated clause.

9. Predications main word. This refers to the predicative complements of the verbs *be*, *appear*, *seem*, *look*, *sound*, *smell*, *taste*, *feel*, *become* and *get*.
10. Closest noun phrase (head) to the left
11. Closest noun phrase (head) to the right
12. Presence of a cleft construction. A binary feature which refers to constructions containing adjectives which trigger extraposed sentential subjects as in '*So it's difficult to attack malaria from inside malarious societies, but it's equally tricky when we try to attack it from outside of those societies.*.
13. Closest adjective to the right
14. VerbNet selectional restrictions of the verb. VerbNet (Kipper et al. 2008) specifies 36 types of argument that verbs can take. We limited ourselves to the values of 'abstract', 'concrete' and 'unknown'.
15. Likelihood of head word taking an event subject (verbs only). An estimate of the likelihood of a verb taking a event subject was computed over the Annotated English Gigaword corpus version 5 (Napoles, Gormley, and Van Durme 2012). We considered two cases where an event subject appears often and may be identified by exploiting the parse annotation of the Gigaword corpus. The first case is when the subject is a gerund and the second case is composed of *this* pronoun subjects.
16. NADA probability. The probability that the non-referential *it* detector, NADA (Bergsma and Yarowsky 2011), assigns to the instance of *it*.

We also experimented with other features and options. For features 2 and 3, a window of three tokens showed a degradation in performance. For features 9 and 10, we experimented with adding their WordNet type (WordNet (Princeton University 2010) contains 26 types of nouns), but this had no effect. The feature combination of noun and adjectives to the left or right also had no effect.

For feature 15, the likelihood of the head verb taking an event subject, we also tried computing it using the ACE corpus (Walker et al. 2005). The ACE corpus is much smaller than the Gigaword corpus, but has been annotated for entities, relations and events. Initially, we thought that the annotation for events would provide reliable statistics on when verbs take an event subject. Although this intuition turned out to be true, it only served

us partially. The extraction from the ACE corpus produced a set of 918 verbs taking an event subject, but these verbs have a journalistic genre nature, since the original data come from journalistic texts. Examples of the extracted events include *to bomb*, *to die*, *to explode*, among others. Our own data, in contrast, comes from TED talks and therefore there was not much overlap between the type of events from both corpora. For instance, from the 501 verbs in the development data, only 309 are found in the set extracted from the ACE corpus, while using the Gigaword corpus, 495 are found. This is no surprise, since the extraction from the Gigaword corpus produced 124,437 verbs taking an event subject. In a dimension that is far off the limits of this dissertation, this example shows a tiny part of the difficulty to infer semantic clues from data.

6.5 Experiment 1: Distinguishing three readings of *it*

All the features listed above were fed into a first MAXENT classifier, obtaining the results listed below in Section 6.4.

A quick scan of Table 6.4 anticipates one of our conclusions: predicting event reference pronouns is a complex problem. Our classifier is more balanced than both baselines, achieving encouraging results with all the classes. However, compared to both of the baselines, the event class shows a small improvement.

A manual inspection of the results shows that discriminating between anaphoric and event reference instances of *it* is indeed an intricate process. Determining the presence or the lack of a specific (nominal) antecedent requires the understanding of the complete coreference chain. Take for instance the following example taken from a dialogue in the corpus:

(2) $_1$You're part of a generation that grew up with the Internet, and it seems as if you become offended at almost a visceral level when you see something done that you think will harm the Internet. $_2$Is there some truth to it? $_3$*it* is. $_4$I think it's very true. $_5$This is not a left or right issue. $_6$Our basic freedoms, and when I say our, I don't just mean Americans, I mean people around the world, *it*'s not a partisan issue.

In the example above (2), the first highlighted *it* is a pronoun with event reference func-

tion while the second is an anaphoric pronoun. With access to the whole coreference chain, one can see that the *it* in sentence 3 refers to the event expressed in the first sentence, therefore it is annotated as an event. This same entity is then referred to with the word *issue* in sentence 5, which in turn becomes the antecedent to the *it* in sentence 6. The classifier, however, labeled these two instances as anaphoric and event respectively.

6.6 Experiment 2: Use of an oracle feature

The results from the previous experiment, such as the case showed in example (2), suggest that the function of a pronoun may be partly determined by their place with respect to other pronouns in the text. With this idea, we looked into the segments from the corpus with more than one instance of *it*. From the 2,031 segments composing the annotated corpus, 349 (17%) contain co-occurrences of between 2 and 7 *it* pronouns within the same segment. In this second experiment, we include the previous *it-label*, when there are several within the same segment, as an additional feature. We consider it an oracle feature, since it is not possible to obtain annotations for new unseen data automatically. Results are reported in Table 6.4.

Gains in performance were obtained using the previous *it-label* as an additional feature. It can be seen in the *w/ oracle feature* section of Table 6.4 that performance improves in the *it-anaphoric* and *it-event* cases but not for the *it-pleonastic* class. This outcome is explained by the fact that both anaphoric and event pronouns are intrinsically referential and therefore potentially part of a bigger coreferential chain including several pronouns. Pleonastic pronouns, on the other hand, are syntactically required but do not corefer.

We tried to approximate the oracle feature by using the relative position of the *it-label* to other *it-labels* within the same sentence (e.g., first, second, etc.). Contrary to the oracle feature, the approximated feature did not lead to any improvement. For both Experiment 1 and Experiment 2, binary classification (event vs. non-event) consistently underperformed when compared to the three class set-up.

MAXENT	14-fold cross-validation			Test-set		
w/o oracle feature	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.627	0.741	0.676	0.716	0.756	0.735
it-pleonastic	0.692	0.565	0.613	0.750	0.726	0.738
it-event	0.579	0.475	0.519	0.564	0.521	0.542
Accuracy	(1328/2506) 0.530			(344/501) 0.687		
w/ oracle feature	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.632	0.750	0.683	0.729	0.785	0.756
it-pleonastic	0.660	0.581	0.607	0.705	0.694	0.699
it-event	0.596	0.502	0.542	0.611	0.538	0.572
Accuracy	(1356/2506) 0.541			(358/501) 0.715		

Table 6.4: Classification results of Experiments 1 and 2 using 14-folds cross-validation and a single test set.

6.7 Contrastive experiments

The results obtained in Experiments 1 and 2 presented an improvement over the *n*-gram baseline, but the improvement over the majority class baseline is much more moderate. In this section, we compare the maximum entropy classifier with other classification algorithms in order to test if better results can be obtained with alternative classification schemes. We tested a support vector machine (SVM) classifier and a simple feed-forward neural network (NN).

The SVM classifier was built using Liblinear (Fan et al. 2008) and trained using a polynomial kernel of degree 4 with parameter C = 10, parameter $\gamma = 0.1$ and bias = 3. Features were scaled to the interval [-1, 1]. Moreover, the second alternative classifier is a feed-forward network built with the Keras package (Chollet 2015). It is a simple network with one hidden layer, optimized with stochastic gradient descend for accuracy and categorical cross-entropy as loss function with a learning rate = 0.001.

The results obtained in these contrastive experiments are given Table 6.5 and it can be noted that they are very similar to those obtained with the MAXENT classifier. Therefore, we suggest that they provide further evidence about the difficulty of the task, in particular identifying the *it-event* class.

	Dev-set			Test-set		
SVM	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.686	0.746	0.715	0.703	0.737	0.720
it-pleonastic	0.846	0.598	0.701	0.724	0.677	0.700
it-event	0.537	0.554	0.545	0.544	0.515	0.529
Accuracy	0.659 (330/501)			0.655 (328/501)		
NN	Precision	Recall	F1	Precision	Recall	F1
it-anaphoric	0.774	0.672	0.720	0.800	0.722	0.759
it-pleonastic	0.543	0.833	0.658	0.677	0.778	0.724
it-event	0.510	0.530	0.519	0.533	0.608	0.568
Accuracy	0.649(325/501)			0.695 (348/501)		

Table 6.5: Classification results of contrastive experiments on development and test sets.

6.8 Self-training experiments

6.8.1 Unlabeled data

Given the small size of the gold-standard data, and with the aim of gaining insight from unstructured and unseen data, we used the MAXENT classifier from Experiment 1 to label additional data from the pronoun prediction shared task at WMT16 (Guillou, Hardmeier, Nakov, et al. 2016). This new *silver-standard* training corpus comprises data from the Europarl (3,752,440 sentences), News (344,805 sentences) and TED talks (380,072 sentences) sections of the shared task training data, and amounts to 1,101,922 sentences.

6.8.2 Recurrent neural network

In the next experiments we present three systems which share an identical architecture but differ in their training data. The architecture used is a deep recurrent neural network (RNN) with word-level embeddings, two layers of Gated Recurrent Units (GRUs) of size 90 and a final softmax layer to make the predictions. The network uses a context window of 50 tokens both to the left and right of the *it* to be predicted. The features described above (Section 6.4) are also fed to the network in the form of one-hot vectors. The system uses the *adam* optimizer and the categorical cross-entropy loss function. We chose this architecture following the example of Luotolahti, Kanerva, and Ginter (2016)[2] and using

[2] `https://github.com/TurkuNLP/smt-pronouns`

the Keras package (Chollet 2015). The first system is trained on the gold-standard data (RNN-GOLD), the second one on the silver-standard data (RNN-NOISY), and the third one on the combination of the two types of data (RNN-COMBINED).

RNN models may be used to capture long range dependencies. Assuming once again that our pronouns of interest are part of a bigger coreference chain, capturing other pronouns and mentions in the same chain, may influence the choice of function label assigned to each of the pronouns.

6.8.3 Results

We report all of the results in Table 6.6. Since they are trained on the same gold-standard data, one would expect RNN-GOLD to perform similarly to MAXENT. However, for the gold-standard data, only the current sentence and two previous sentences were extracted, not the full documents. This strategy, together with the small size of the corpus, does not fully exploit the strengths of this type of model. On the other hand, this may explain the relatively good performance of the RNN-NOISY system. Although the system is trained on noisy data, it has access to full documents.

As expected, RNN-COMBINED performs better than RNN-GOLD and RNN-NOISY. Although it does not perform overwhelmingly better than MAXENT, there are gains in the precision for it-anaphoric class, and in recall for it-pleonastic and it-event classes, suggesting that the system benefits from the inclusion of gold-standard data.

While both RNN-COMBINED and MAXENT are more balanced than the baseline, confusion is observed in the prediction of the it-event and it-anaphoric classes: MAXENT predicts more anaphoric instances as event and the RNN-COMBINED does the opposite, predicting more event instances as anaphoric. An it-event pronoun could be seen as referential pronoun lacking a nominal antecedent. In this sense, the MAXENT classifier identifies more antecedents than there actually are, while the RNN-COMBINED fails to recognize them.

Manual inspection of the results confirmed that MAXENT makes fewer errors in identifying the it-anaphoric class when there is a clear antecedent nearby (49/118 vs. 96/118 errors), as in example (3). RNN-COMBINED, on the other hand, performs better for the it-event class, particularly in difficult cases where the reader must infer the event from the text (12/25 vs. 22/25 errors). This is the case in example (4), where the pronoun *it*

	Dev-set				Test-set			
MAXENT	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.685	**0.758**	**0.719**	(326/501)	0.716	**0.756**	**0.735**	(344/501)
it-pleonastic	**0.884**	**0.543**	**0.633**	0.651	**0.750**	0.726	**0.738**	0.687
it-event	**0.545**	0.541	**0.543**		**0.564**	0.521	0.542	
RNN-GOLD	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.544	0.560	0.552	(221/501)	0.595	0.659	0.626	(250/501)
it-pleonastic	0.274	0.217	0.242	0.441	0.177	0.177	0.177	0.499
it-event	0.355	0.382	0.368		0.436	0.361	0.394	
RNN-NOISY	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.661	0.611	0.635	(286/501)	0.706	0.552	0.620	(286/501)
it-pleonastic	0.725	0.402	0.517	0.571	0.542	0.516	0.529	0.571
it-event	0.438	0.605	0.508		0.455	0.621	0.525	
RNN-COMBINED	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	**0.697**	0.492	0.577	(280/501)	**0.794**	0.530	0.636	(315/501)
it-pleonastic	0.633	**0.543**	0.585	0.559	0.582	**0.742**	0.652	0.629
it-event	0.434	**0.675**	0.529		0.520	**0.746**	**0.613**	

Table 6.6: Comparison of results of the self-training experiments. MAXENT refers to the maximum entropy classifier trained on the gold-standard data only (ParCor corpus and DiscoMT2015.test). RNN-GOLD is the RNN trained on the gold-standard data as well. RNN-NOISY is trained on the silver-standard data (annotated using the MAXENT classifier). RNN-COMBINED is trained on both the silver-standard and gold-standard data.

refers to the *rejection* of mandatory retirement age by people 15 and over.

(3) We trust that they will spend their time with us, that they will play by the same rules, value the same goal, they'll stay with *the game* until *it*'s over.

(4) According to the same survey, 61% of the population aged 15 and over believe that people should be able to continue to work beyond any statutory retirement age: in other words, they reject the idea of a mandatory retirement age. *It* is an important and welcome development.

The inter-annotator agreement kappa score of 0.81 reported for the gold-standard data represents a reasonable upper bound for system performance. According to Guillou (2016, p.81), 18 of the 27 disagreements arose due to differences in opinion as to whether an instance of *it* is used for nominal anaphoric or event reference. In our own evaluation, we counted 35 such ambiguous cases, and found that RNN-COMBINED yields marginally fewer errors than MAXENT (23 vs 27).

Examining these ambiguous cases more closely, we looked at the intersection of errors between MAXENT and RNN-COMBINED, focusing on the cases where both systems disagreed with the gold labels. We found that in 11 of 72 cases, the systems prediction was better than the gold label, pointing to a possible 15% error rate in the gold-standard annotation.

6.9 Source-aware language model

In this section we include the *it* disambiguation task in the task of cross-lingual pronoun prediction. In order to predict pronoun translations, we used the classifier described in Experiment 1 as part of an *n*-gram language model trained over target lemmas. In addition to the pure target lemma context, our model also has access to the identity of the source language pronoun, which, in the absence of number inflection on the target words, provides valuable information about the number marking of the pronouns in the source and opens a way to inject the output of the pronoun type classifier into the system.

SOURCE:	*It* 's got these fishing lures on the bottom .
TARGET LEMMAS:	REPLACE_0 avoir ce leurre de pêche au-dessous .
SOLUTION:	*ils*
LM TRAINING DATA:	It REPLACE_*ils* avoir ce leurre de pêche au-dessous .
LM TEST DATA:	It REPLACE avoir ce leurre de pêche au-dessous .

Figure 6.3: Data for the source-aware language model.

6.9.1 Design

Our source-aware language model is an *n*-gram model trained on an artificial corpus generated from the target lemmas of the 2016 shared task training data (Figure 6.3). To every REPLACE tag occurring in the data, we insert the source pronoun aligned to the tag (without lowercasing or any other processing). The alignment information attached to the REPLACE tag in the shared task data files is stripped off. In the training data, we instead add the pronoun class to be predicted. Note that all REPLACE tags are placeholders for one word translations guaranteed to correspond to a source pronoun *it* or *they* according to the shared task data preparation (Hardmeier, Nakov, et al. 2015; Guillou, Hardmeier, Nakov, et al. 2016). The *n*-gram model used for this component is a 6-gram model with modified Kneser-Ney smoothing (S. Chen and Goodman 1998) trained with the KenLM toolkit (Heafield 2011).

To predict classes for an unseen test set, we first convert it to a format matching that of the training data, but with a uniform, unannotated REPLACE tag used for all classes. We then recover the tag annotated with the correct solution using the `disambig` tool of the SRILM language modelling toolkit (Stolcke et al. 2011). This tool runs the Viterbi algorithm to select the most probable mapping of each token from among a set of possible alternatives. The map used for this task trivially maps all tokens to themselves with the exception of the REPLACE tags, which are mapped to the set of annotated REPLACE tags found in the training data.

6.9.2 English-French *it* disambiguation language model

We used the classifier described in Section 6.2 to annotate all instances of *it* from the source side of the data which were mapped to a REPLACE item according to the alignment provided. Afterwards, a new source-aware language model is trained in the manner

described in Section 6.9.1. In this way, instead of the sentence *'It 's got these fishing lures on the bottom .'* presented in Figure 6.3, the system receives the labeled input *'it-anaphoric 's got these fishing lures on the bottom .'* All the data provided for the shared task was used in training this system.

6.9.3 Results

The macro-averaged recall (official metric of the WMT 2016 shared task) obtained is 57.03%. This is slightly lower than the score of 59.84% which was obtained by the system without the *it* labels (Table 6.7). However, some pronouns present better scores using the system with the it-labels than the system without them. Precision, in particular, is higher. This outcome is expected for the pronoun *cela*, which is the French neuter demonstrative pronoun frequently used for event reference. However, there are also gains in precision for *on*, *elles* and *ils* while maintaining recall.

w/o labels Macro R: 59.84%			
Pronoun	Precision	Recall	F1
ce	**89.66**	76.47	82.54
elle	**40.00**	60.87	48.28
elles	27.27	12.00	16.67
il	**63.24**	70.49	66.67
ils	67.82	83.10	74.68
cela	76.47	41.94	54.17
on	36.36	44.44	40.00
OTHER	**88.37**	89.41	88.89
w/ labels Macro R: 57.03%			
Pronoun	Precision	Recall	F1
ce	89.09	72.06	79.67
elle	31.25	43.48	36.36
elles	**30.77**	16.00	21.05
il	54.43	70.49	61.43
ils	**69.41**	83.10	75.64
cela	**86.67**	41.94	56.52
on	**40.00**	44.44	42.11
OTHER	85.71	84.71	85.21

Table 6.7: Source-aware language model with and without *it* disambiguation labels.

SOURCE:	*it-anaphoric* just takes a picture of objective reality as *it-anaphoric* is .
LM W/O LABELS:	il OTHER
LM W/ LABELS:	elle OTHER
BASELINE:	cela OTHER
REFERENCE	*elle* prendre juste un image objectif de *la* réalité .

Figure 6.4: Examples of predictions of the final systems. The reference translation is lemmatized.

6.9.4 Manual analysis

In order to further investigate the impact of the disambiguation of *it* on the prediction task, we isolated the cases where the French pronouns are translations of *it*. We relied on the alignment information from the shared task data to separate the French translations of *it* and *they*. A gold subset of 233 *it* cases was obtained. Precision, recall and F-score were then computed in the usual manner (Table 6.8).

This second evaluation shows that the improvements obtained for *cela* and *on* are legitimately due to the *it* disambiguation labels. While other classes do not show the same gain in performance, a manual analysis reveals somewhat fewer incoherence errors. For instance, the system with the labels consistently misclassifies *elle* as *il*, a coreference error not addressed by the it-anaphoric label. Besides, with the exception of the *il* class, the marginal distributions (total column and row) of the system with the labels are closer to each other than that of the system without the labels.

Looking at the predictions, we confirmed that both source-aware language models produced identical results almost all of the time, while the system without the labels produces more correct predictions in total. However, there are some interesting examples where the system with the labels outperforms both the baseline and the un-labeled one. A contrastive example can be seen in Figure 6.4.

6.10 Conclusion

Distinguishing between anaphoric and event reference realisations of *it*, as it turns out, is a very complex task. In particular, it can be difficult to determine the antecedent of an event reference pronoun. event reference is a subtype of anaphoric reference after all.

Shared task LM baseline

	Classified as →					cela/					
Gold ↓	OTHER	on	elle	ce	il	ça	elles	ils	Total	P	R	F1
OTHER	38	1	2	8	4	1	0	0	54	80.85	70.37	75.25
on	0	4	0	0	0	0	0	0	4	26.67	**100**	42.11
elle	1	0	11	2	6	2	0	0	22	**57.89**	50.00	**53.66**
ce	3	0	0	50	8	1	0	0	62	76.92	**80.65**	78.74
il	1	7	2	4	41	3	0	0	58	62.12	70.69	66.13
cela/ça	3	3	3	1	7	14	0	0	31	66.67	**45.16**	53.85
elles	1	0	1	0	0	0	0	0	2	0	0	0
ils	0	0	0	0	0	0	0	0	0	0	0	0
Total	47	15	19	65	66	21	0	0	233			

Source-aware LM without it-labels

	Classified as →					cela/					
Gold ↓	OTHER	on	elle	ce	il	ça	elles	ils	Total	P	R	F1
OTHER	48	0	2	1	3	0	0	0	54	**85.71**	**88.89**	**87.27**
on	0	3	0	0	1	0	0	0	4	33.33	75.00	46.15
elle	0	0	14	2	5	0	0	1	22	43.75	**63.64**	51.85
ce	4	0	2	49	5	2	0	0	62	**90.74**	79.03	**84.48**
il	1	4	8	2	41	2	0	0	58	**64.06**	70.69	**67.21**
cela/ça	2	2	5	0	9	13	0	0	31	76.47	41.94	54.17
elles	1	0	1	0	0	0	0	0	2	0	0	0
ils	0	0	0	0	0	0	0	0	0	0	0	0
Total	56	9	32	54	64	17	0	1	233			

Source-aware LM with it-labels

	Classified as →					cela/					
Gold ↓	OTHER	on	elle	ce	il	ça	elles	ils	Total	P	R	F1
OTHER	45	1	1	1	6	0	0	0	54	83.33	83.33	83.33
on	0	3	0	0	1	0	0	0	4	**37.50**	75.00	**50.00**
elle	0	0	10	1	11	0	0	0	22	33.33	45.45	38.46
ce	4	0	4	45	8	1	0	0	62	90.00	72.58	80.36
il	2	2	10	2	41	1	0	0	58	54.67	70.69	61.65
cela/ça	3	2	5	1	7	13	0	0	31	**86.67**	41.94	**56.52**
elles	0	0	0	0	1	0	0	1	2	0	0	0
ils	0	0	0	0	0	0	0	0	0	0	0	0
Total	54	8	30	50	75	15	0	1	233			

Table 6.8: Comparison of the source-aware language models with *it* disambiguation labels evaluated on translations of *it* only. Note that this gold subset does not include the class *ils*. However, since this is a subset, some errors appear which were learned from the *they* training examples.

Even though the results obtained with the MAXENT classifier are much better than the results of the *n*-gram baseline, they remain modest.

The self-training experiments demonstrated the benefit of combining the gold-standard data with noisy data labeled by the MAXENT classifier. Yet, it would seem that more gold data is needed for a bigger impact. In addition, we found that the RNN-COMBINED system is better at handling difficult and ambiguous referring relationships, while the MAXENT performed better at identifying patterns, such as those common to the it-pleonastic class.

As the results suggest, the task presented in this chapter can be beneficial for correct cross-lingual pronoun prediction and coreference resolution. However, there is definitely room for improvement. In this sense, further (error) analysis focused on understanding the differences between nominal and event reference seems a necessary next step. With this chapter we conclude our work in pronominal anaphora. In the following two chapters, we shift our focus of attention from pronominal reference and pronoun translation to verbal reference and verbal tenses translation.

Chapter 7

Tense translation modeling: exploiting grammatical tense

7.1 Introduction

While in previous parts of this work we have focused on pronominal reference and pronoun translation, in the next two chapters we treat verbal reference and verbal tenses translation. In addition, the experiments developed in this part are based on full phrase-based statistical machine translation (SMT).

Verbal tenses are the primary linguistic source of temporal reference, i.e. the localization of events and states in time. Verbal tenses express a relative temporal location of events and states[1] with respect to the moment of speech and with respect to each other, and not only relative to a fixed linear time continuum. Because tenses require a previously established temporal referent, they are considered anaphoric. For instance, the present tense has the moment of speech as referent, while the past tenses may have a temporal adjunct, such as *minutes before*, *at five o'clock yesterday*, *when I woke up*, or *earlier in the day* (Moens and Steedman 1988). For example, the past perfect form *had been* in sentence (1) (taken from Moens and Steedman (1988)[2]) locates the event of Bonadea's love affair before the moment of speech, *two weeks later*.

[1] In computational linguistics approaches the term *event* is often used as a synonym of the more linguistic terms *eventuality* or *situation*, which are generic concepts including all aspectual types of verbs. In the linguistic literature, however, the term *event* does not include verbs considered as *states*.

[2] Extracted from *Der Mann ohne Eigenschaften* by Robert Musil.

(1) Two weeks later, Bonadea had already been his lover for a fortnight.

The analysis of tenses as anaphoric has its roots in Reichenbach (1947)'s formalization of the reference time. He proposed three temporal points, event point E, moment of speech S and reference point R, and two temporal relations, precedence and simultaneity, to account for temporal reference. In this framework, a verbal form such as the past perfect receives the following formalization: $E < R < S$. Hence the meaning of the form 'past perfect' indicates that the moment when the event took place is previous to the reference moment which is previous to the speech moment (Loáiciga and Grisot 2016).

However, contrary to pronominal anaphora, in which anaphors have a unique antecedent, a tense may refer to changing points in time. Moens and Steedman (1988) give the following example, '*At exactly five o'clock, Harry walked in, sat down, and took off his boots.*', in which they argue, the reference point moves away from its starting point *five o'clock*'. Since a different order like '*At exactly five o'clock, Harry took of his boots, sat down and walked in.*' would change the logical interpretation of events, all the mentioned events do not have the same reference point.

This view of verbal tenses finding their temporal meaning in the sentence they are contained in will be further discussed in Chapter 8. In the present chapter, we will concentrate on the concept of grammatical tense. Tense, in this sense, is defined morphologically, it refers to particular verbal forms, to the pairing of a forms with meanings (Declerck 2007, pp. 52-53). For instance, even in a morphologically relatively poor language such as English, we know that a -ed verb ending, in the absence of future or conditional auxiliaries, often indicates past tense or passive.

In spite of the vast amount of linguistic literature on the analysis of verbal tenses, there is much less work in the context of machine translation and even less in the context of SMT. However, combining SMT systems with linguistic insights on verbal tenses can contribute to a text's cohesion and coherence, as Ye, Schneider, and Abney (2007, p. 521) assert,

> Correct translation of tense and aspect is crucial for translation quality, not only because tense and aspect are relevant for all sentences, but also because temporal information is essential for a correct delivery of the meaning of a sentence.

In this chapter, we annotate a parallel English-French corpus automatically with tense

labels. We then look at the distribution of verbal tenses between English and French and analyze the translation correspondences between the two. Last, we present a machine translation experiment using the tense labels to improve the translation of English tenses to French.

7.2 Automatic annotation of VPs with tense and voice

Motivated by Grisot and Cartoni (2012)'s study about the translation of verbal tenses between English and French, we perform a similar study scaling up the data and using automatic methods. Grisot and Cartoni (2012) examined a translation divergency when translating the English simple past tense into French. They considered three potential tense translations in French: *passé composé* (PC), *imparfait* (IMP) and *passé simple* (PS). That study, however, was completed manually using a very small corpus.

We work with the English-French portion of the Europarl corpus of European Parliament debates (Koehn 2005). The verb phrases (VPs) in the corpus were identified and annotated automatically with tense and voice information. We consider VPs narrowly: only verbal elements such as the head and auxiliaries and participles are retained, and not its internal arguments or adjuncts as in a constituency grammar definition. The annotation process was divided in two main parts: a) word-level alignment and morpho-syntactical analysis, and b) tense annotation.

7.2.1 Alignment and morpho-syntactical analysis

The English side of the parallel corpus is parsed with a dependency parser (Henderson et al. 2008) and the French side is parsed with Morfette (Chrupała, Dinu, and van Genabith 2008), which produces lemmas and morphological tags. From the parsing of the English sentences we retain the position, POS tags, heads and the dependency relation information. For the French side, we use both the morphological tags and the lemmas produced by Morfette. The parsing information and the corpus are then aligned at the word-level using bidirectional alignments produced with GIZA++ (Och and Ney 2003).

As an alternative to Morfette for the French analysis, we also experimented using the Lefff Lexicon (Sagot 2010). However, this resource proved to be inadequate for our purposes, since homographic finite verbal forms are not disambiguated. For example,

the first and third person singular of verbs ending in *-er* in French have identical forms for the *indicative* and *subjunctive* moods in the present tense, such that, a form as *parle* ('talk') is simultaneously identified as *present-indicative*, *present-subjunctive* and even second person *imperative*.

Position	**EN token**	**POS-tag**	**Head**	**Dependency Relation**	**FR token**	**Lemma**	**Morpho-syntactical information**
1	Thank	VB	8	DEP	Merci	merci	N_C-ms
2	you	PRP	1	OBJ			
3	,	,	8	P	,	,	PONCT_W
4	Mr	NNP	5	NAME	Monsieur	monsieur	N_P-ms
5	Segni	NNP	8	ADV	Segni	segni	N_P-ms
6	,	,	8	P	,	,	PONCT_W
7	I	PRP	8	SBJ	je	je	CL_suj-1ms
8	shall	MD	0	ROOT	ferai	faire	V-indicatif-present1s
9	do	VB	8	VC			
10	so	RB	9	ADV			
11	gladly	RB	10	AMOD	bien volon-tiers	bien volon-tiers	ADV ADV
12	.	.	0	ROOT	.	.	PONCT_S

Figure 7.1: Example of alignment between the English-French parallel corpus and the parsing information of each side.

7.2.2 Determining and labeling tense

The second processing stage consists in a set of hand-written rules used to identify VPs and assign them a tense label using the information illustrated in Figure 7.1.

The English tense is determined first. For this, potential verbs are recognized on the basis of their POS tags (MD, VB, VBN, VBD, VBG, VBP, VBZ and RP)[3]. Then, by checking their heads and dependency relations, it is established if they are either single or compound forms. For example, if two words tagged as MD (Modal) and VB (Verb Base-form) are found, 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

[3]The abbreviations stand as follows: MD-modal, VB-verb based form, VBN-past participle, VBD-past tense, VBG-verb -ing form, VBP-non-3rd person singular present, VBZ-3rd person singular present and RP-particle.

VB) is interpreted as a valid VP. In this particular case, the first word is further tested in order to disambiguate between a future (i.e., *will*, *shall*) or a conditional (i.e., *should*, *would*, *ought*, *can*, *could*, *may*, *might*) verb. Once all VPs in a sentence are identified, each one is evaluated in order to assign tense and voice labels. We have defined a set of heuristics to assign the labels based on the patterns of the POS-tags of the tokens composing the VP (cf. Appendix A).

The French tense is determined after the English tense. Unlike pronouns, we assume that English VPs are translated mostly by French VPs as well. Therefore, we identify French VPs by the tokens aligned with the already identified English VPs. The morpho-syntactical information and lemmas associated to each of the words constituting the French VP are concatenated and matched against our heuristics, in an identical process to the one on the English side.

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'). For instance, while all forms of passive voice use the auxiliary *être* ('to be'), only a small set of intransitive verbs use it in their compound forms. In this sense, relying on lemmas was essential for French tense identification.

Since manually annotated data is not only expensive but time-consuming to produce, creating data with automatic tools is advantageous to generate resources. However, the accuracy of NLP tools is imperfect and their errors sum up when used in a pipeline. In our case, we had seen some trouble with the analysis of long-distance dependencies by the parser, as illustrated by *will be examined* in example (2). These are constructions where the VP compound is split for a variable length of tokens, making it difficult for the parser to connect the parts.

(2) It *will*, I hope, *be examined* in a positive light.

We detected a similar problem with the word-alignments. Some particle verbs and verbal tenses with multiple auxiliaries are only partially aligned. Take for instance (3), where the VP *have done* is not aligned with the French auxiliary, but only with the past participle *fait* (4).

(3) a. EN. I regret this since we are having to take action because others *have not*

done their jobs .

b. FR. Je le déplore car nous devons agir du fait que que d'autres *n'ont pas fait* leur travail .

(4) have not done .
n'pas fait .

7.2.3 Tense translation between English and French

A total of 423,235 sentences from the corpus were labeled. From this, 3,816 sentences were discarded due to mistmaches between the parser, Giza++ and Morfette outputs, leaving us with 419,420 sentences.

There are a total of 673,844 total English VPs in the corpus: 454,890 finite (67.51%) and 218,954 non-finite (32.49%). Non-finite forms correspond to infinitives, gerunds and past participles acting as adjectives. Since our interest is strictly on verb tenses, non-finite forms are not further examined.

For each English VP with a tense label, we considered whether the French side label was acceptable according the rules. Since the French tenses are determined on the basis of the English VP identification process and on the alignment information, the possibility of erroneous labels is bigger. Table 7.1 shows the number of VPs for each English tense label, as well as the number of pairs with an acceptable label on the French side. The first column shows the identified tenses in English. Even though the tense status of the *future* and the *conditional* forms is at least debatable, we extracted them for description purposes. The second column shows the distribution of each tense in French. The overwhelming disparity between the quantity of *present* tense and all of the other tenses is to be noted: this tense alone represents ca 60% of the total finite VPs.

On average about 81% of the pairs are selected at this stage. Overall, our method thus preserves slightly more than half of the input VP pairs (54.7%[4]), but ensures that both sides of the verb pair have acceptable labels. To estimate the precision of the annotation (and noting that the above figure illustrates its "recall" rate), we evaluated manually a set of 413 VP pairs sampled from the final set, in terms of the accuracy of the VP boundaries and of the VP labels on each side. The results are presented in Table 7.2. In summary, almost 90% of VP pairs have correct English and French labels, although not all of them

[4] $67.51\% \times 81\%$

English tense	EN tense labels		FR tense labels	
Present	270 145	59.4%	219 489	59.8%
Present perfect	49 041	18%	43 433	11.8%
Present continuous	22 364	4.9%	19 118	5.2%
Present perfect continuous	1 104	0.2%	979	0.3%
Simple past	52 198	11.5%	39 475	10.8%
Past perfect	1 898	0.4%	1 520	0.4%
Past continuous	1 135	0.2%	878	0.2%
Past perfect continuous	31	0.0%	26	0.0%
Future	17 743	03.9%	12 963	3.5%
Future perfect	167	0.0%	133	0.0%
Future continuous	675	08.1%	546	0.1%
Future perfect continuous	1	00.0%	1	0.0%
Conditional	36 702	08.1%	27 189	7.4%
Conditional perfect	1258	0.3%	1 059	0.3%
Conditional continuous	415	0.1%	322	0.1%
Conditional perfect continuous	8	0.0%	7	0.0%
Total	454 890	100%	367 138	100%

Table 7.1: Number of annotated finite VPs for each tense category in 419,420 sentences from Europarl.

	VP boundaries		Tense labels	
	English	French	English	French
Correct	97%	80%	95%	87%
Incorrect	1%	4%	5%	13%
Partial	2%	16%	–	–

Table 7.2: Human evaluation of the identification of VP boundaries and of tense labeling over 413 VP pairs.

have perfect VP boundaries.

Since in this work we chose to focus on the English simple past translation into French, we concentrate now on the subset of the data corresponding to the present and past tenses. This subset amounts to a total of 322 086 finite VPs, 70.81% of the total shown in Table 7.1. Table 7.3 and Table 7.4 contain the verbal tenses translation figures of this subset given as number of occurrences and percentage respectively.

The distribution of tenses between the two languages reveals that tense translation, as pronouns, is not one-to-one. All of the past tenses present several translation possibilities. The English SP, in particular, can be translated mainly as *passé composé* (PC),

French	English								
	Past continous	Past perfect continous	Past perfect	Present continuous	Present perfect continuous	Present perfect	Present	Simple Past	**Total**
Imparfait	462	7	**365**	146	18	463	1,510	**8,060**	11,031
Impératif	–	–	–	37	1	6	203	11	258
Passé composé	139	2	**214**	282	325	**26,521**	**1253**	**19,402**	48,138
Passé récent	–	–	1	8	3	187	2	3	204
Passé simple	4	–	6	16	2	54	42	374	498
Plus-que-parfait	27	8	**782**	2	4	217	22	1,128	2,190
Présent	216	9	102	18,077	617	**14,736**	211,334	**9,779**	254,870
Subjonctif	15	–	28	258	6	**1,053**	2,969	568	4,897
Total	863	26	1,498	18,826	976	43,237	217,335	39,325	322,086

Table 7.3: Raw counts of the distribution of the translation labels of 322,086 VPs in 203,140 sentences.

French	English								
	Past continous	Past perfect continous	Past perfect	Present continuous	Present perfect continuous	Present perfect	Present	Simple Past	**Total**
Imparfait	53.53	26.92	**24.37**	0.78	1.84	1.07	0.69	**20.50**	3.42
Impératif	–	–	–	0.20	0.10	0.01	0.09	0.03	0.08
Passé composé	16.11	7.69	**14.29**	1.50	33.30	**61.34**	0.58	**49.34**	14.95
Passé récent	–	–	0.07	0.04	0.31	0.43	0.00	0.01	0.06
Passé simple	0.46	–	0.40	0.08	0.20	0.12	0.02	0.95	0.15
Plus-que-parfait	3.13	30.77	**52.20**	0.01	0.41	0.50	0.01	2.87	0.68
Présent	25.03	34.62	6.81	96.02	63.22	**34.08**	97.24	**24.87**	79.13
Subjonctif	1.74	–	1.87	1.37	0.61	**2.44**	1.37	1.44	1.52
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 7.4: Distribution (%) of the translation labels for 322,086 VPs in 203,140 sentences. A zero indicates fewer than 1% of occurrences, while blanks correspond to instances that did not occur at all.

imparfait (IMP), *passé simple* (PS) and also as *présent* (PRES), entailing then a significant translation ambiguity ($p < 0.001$, χ (7, N=39,325)=70287.5). The PC translation, however, comprises almost 50% of the total translations. The present perfect is another example. This tense can be translated either as *passé composé* (61.34%), *présent* (34.08%) or *subjonctif* (2.44%) mainly. The *subjonctif* or subjunctive is rather a mood

than a tense. However, since we observed many of these particular verbal forms we decided to include it in this description of the translation possibilities of the English past tenses into French. The choice of indicative or subjunctive mood is strongly dependent on the selectional preferences of the verb. For instance, predicates of 'desire', 'uncertainty' or 'probability', 'directives' and 'causatives' might trigger the subjunctive mood in French and Spanish (Ahern 2005). Mood is certainly an important factor for the interpretation of the sentential context a verb appears in. However, the analysis of this property is beyond the scope of this thesis.

The SP figures confirm the translation ambiguity reported by Grisot and Cartoni (2012) and present new ones. Besides, while Grisot and Cartoni (2012) reported a 1 to 3 ambiguity for the English SP, we found that the *présent* is a translation possibility quite frequent as well.[5].

7.3 Tense-aware Statistical Machine Translation System

Aiming at assessing the usefulness of the grammatical tense to improve the automatic translation of the English SP into French, we have built two systems. The first is a tense-aware system trained on the automatically annotated corpus just described. The second is a baseline system trained on an unannotated version of the same corpus. For the tense-aware system, we use the French tense labels as disambiguation labels on the English SP verbs. In this sense, the tense-aware system serves as an oracle system since we have the French tense annotation at testing time, which would not be the case in a machine translation scenario.

7.3.1 Design

We used phrase-based translation models built with the Moses toolkit (Koehn, Hoang, et al. 2007) for the two systems. In addition, for the tense-aware system, we used factored translation models (Koehn and Hoang 2007), which allow to integrate arbitrary linguistic markup (i.e., factors) at the word level. In our case, the SP verbs in the English sentences receive one French tense label (e.g. 'was|IMP' for *imparfait*), and all

[5]The annotated corpus can be downloaded at `https://www.idiap.ch/dataset/tense-annotation`

other words are set to the '|null' factor. The two systems were built by partitioning the total 203,140 annotated sentences as follows: 196,140 sentences for training; 4,000 sentences for tuning; and 3,000 sentences for testing. A 5-gram language model trained on the entire French side of Europarl version 5 with SRILM (Stolcke et al. 2011) is used. Optimization weights were fixed using Minimum Error Rate Training (MERT) (Och 2003).

7.3.2 Results and discussion

We evaluated the 3,000 sentences of the test-set with the BLEU (Papineni et al. 2002) and METEOR (Denkowski and Lavie 2011) automatic metrics and using manual error inspection as well. The scores obtained are given in Table 7.5. It can be noted that the tense-aware system gained 0.5 points over the baseline for the BLEU score. The METEOR score shows a positive difference of 0.0029 points between the two systems. Since this score is calculated not only on the basis of exact matches but also on stems, the small difference means that only few verb stems are changed. This is expected since a tense-aware system should mainly modify inflectional suffixes, but not the stems.

System	BLEU	METEOR
Baseline	27.67	0.4912
Tense-aware	28.17	0.4941

Table 7.5: BLEU and METEOR scores obtained by the baseline and tense-aware systems.

	Baseline	Tense-aware	# sentences
Imparfait	24.10	25.32	122
Passé composé	29.80	30.82	359
Impératif	19.08	19.72	4
Passé simple	**13.34**	**16.15**	6
Plus-que-parfait	**21.27**	**23.44**	17
Présent	27.55	27.97	2618
Subjonctif	26.81	27.72	78
Passé recent	**24.54**	**30.50**	3

Table 7.6: BLEU scores per expected French tense for the three systems. Largest score increases are boldfaced. The number of sentences for each class is given in the last column.

The increment of BLEU is important, as the detailed BLEU scores per tense presented

in Table 7.6 reveal. Indeed, when each expected French tense is observed in detail, the least frequent tenses seem to obtain the biggest improvements in translation quality. The tense-aware system obtained improved results throughout all the tenses, but the *passé simple, plus-que-parfait* and *passé récent* display particularly better results than the baseline. These figures suggest that high-frequency tenses such as the present tense – which do not have virtually any translation ambiguity, as evidenced by the 97.24% of this tense translated as French *Présent* tense– tend to hide in the overall scores the genuine improvement of the tense-aware system on ambiguous tenses.

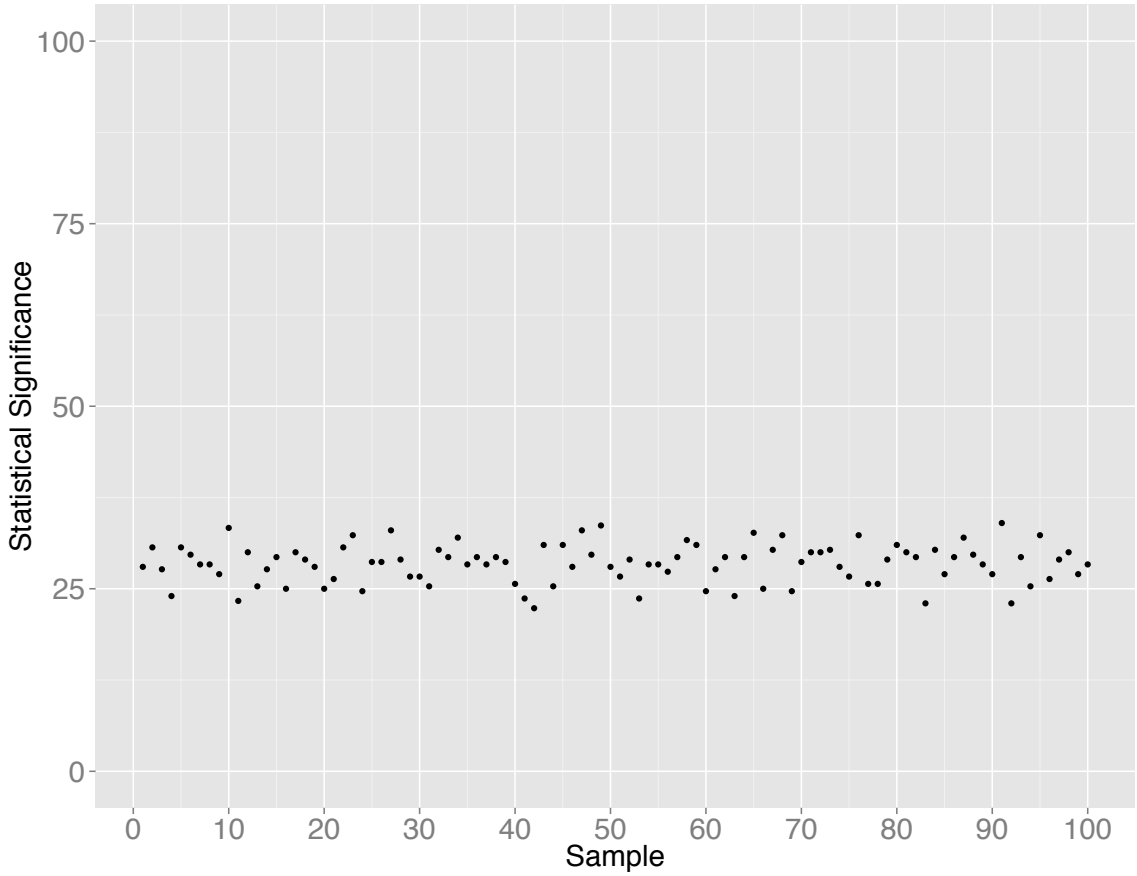


Figure 7.2: Results of paired bootstrap with resampling test. The x axis shows the ID of the samples and the y axis displays the percentage of sentences per sample which obtained a higher BLEU score than the sentences in the baseline.

Since automatic MT scores can be difficult to interpret, a bootstrap resampling significance test as introduced by Koehn (2004) was completed. This test estimates the difference in performance of one SMT system in comparison to another. Using the test set,

100 paired samples of 300 sentences each containing at least one verb in the SP tense are generated. The BLEU score is computed and recorded for each sentence in each sample. Results are depicted in Figure 7.2. It can be seen that ca 27% of the sentences translated by the tense-aware system are constantly scored higher than the sentences translated by the baseline system. A 80% confidence level estimate computed over another sampling of 100 samples of 300 sentences places the confidence interval of the differences in BLEU scores between the systems at [0.1, 0.8] (Zhang, Vogel, and Waibel 2004).

A qualitative assessment of the systems was done by means of a detailed manual evaluation of 313 sentences, comprising 654 VPs from the test set. The results are shown in Table 7.7. Each annotated VP was evaluated in three different categories. **TAM** refers to the tense, aspect and mode features. With **lexical choice** we assess the correctness of the verbal lemma. This criterion captures the cases in which the TAM features of a VP were improved but the lemma itself changed, being then penalized by BLEU. Finally, **agreement** refers to whether a translation is free from errors of person and number agreement. For the first two categories, we evaluated if the translation was different than the reference yet correct ($\neq$ ref) or identical ($=$ ref).

System	**TAM**			**Lexical choice**			**Agreement**		Total VPs
	Incorrect	Correct		Incorrect	Correct		Incorrect	Correct	
		$\neq$ ref.	$=$ ref.		$\neq$ ref.	$=$ ref.			
Baseline	206	61	387	47	267	340	118	536	654
	32%	9%	59%	7%	41%	51%	18%	82%	100%
Tense-aware	52	39	563	60	247	347	122	532	654
	8%	6%	86%	9%	38%	53%	19%	81%	100%

Table 7.7: General results of the manual evaluation of 313 sentences from the test set.

In terms of tense translation the tense-aware model outperformed the baseline by an average of 24% and up to 27%. This is congruent with the results of the bootstrap test (Figure 7.2). Concerning the lexical choice and the agreement categories, they did not change much between the three systems. When looking at the results per French translated tense (Table 7.8) we confirmed that low-frequency verbs are better translated by the tense-aware system, for instance the *passé simple* and the *passé récent*.

On the other hand, the results for the *imparfait* and the *subjonctif* tenses (boldfaced in Table 7.8) reveal that English tenses with a real translation ambiguity were better translated by the tense aware system. For instance, while most of the *present perfect* English VPs were translated as *passé composé* by the baseline – since this is the most frequent translation with up to 61% of the instances (Table 7.4), the tense aware models

boosted the instantiation of the *imparfait* tense in French.

French tense	System	TAM			Lexical Choice			Agreement		Total VPs
		Incorrect	Correct		Incorrect	Correct		Incorrect	Correct	
			≠ ref	= ref		≠ ref	= ref			
Imparfait	Baseline	**82**	**15**	**41**	7	56	75	27	111	138
	Tense-aware	**13**	**4**	**121**	14	51	73	27	111	
Passé composé	Baseline	28	6	129	14	68	81	32	131	163
	Tense-aware	14	5	144	8	66	89	32	131	
Passé récent	Baseline	2	1	0	0	2	1	0	3	3
	Tense-aware	0	0	3	2	0	1	2	1	
Passé simple	Baseline	3	2	1	0	3	3	0	6	6
	Tense-aware	0	1	5	2	2	2	2	4	
Impératif	Baseline	1	2	1	0	3	1	2	2	4
	Tense-aware	0	1	3	0	3	1	1	3	
Plus-que-parfait	Baseline	6	4	8	0	7	11	7	11	18
	Tense-aware	2	2	14	1	9	8	8	10	
Présent	Baseline	21	20	201	16	93	133	34	208	242
	Tense-aware	12	19	211	13	87	142	25	217	
Subjonctif	Baseline	**63**	**11**	**6**	10	35	35	16	64	80
	Tense-aware	**11**	**7**	**62**	20	29	31	24	56	

Table 7.8: Results per expected tense of the manual evaluation. Only the most frequent tenses were evaluated.

In Figure 7.3 we present an example taken from the test set. The first verb is incorrectly translated with a French *infinitif* by the baseline system, but correctly by the tense-aware system. The second verb is also incorrectly translated into a *présent*, indicative mode, while a *subjonctif* was required. Some of the surrounding words are of variable correctness, since the clause '*that is better at showing how much money is going into the EU.*' seem hard for the systems.

SOURCE	We want particularly to emphasise that we support a system that we *support* a system that *is* clearer than the current one and that is better at showing how much money is going into the EU.
BASELINE	Nous voulons, en particulier, de souligner que nous *soutenir* un système qui *est* plus claire que le système actuel et que est mieux à la façon dont beaucoup argent dans l'UE.
TENSE-AWARE	Nous voulons, en particulier, de souligner que nous *soutenons* un système qui *soit* plus clair que ce que le programme actuel est mieux à la façon dont beaucoup argent dans l'UE.
REFERENCE	Nous tenons à insister tout particulièrement sur le fait que nous *soutenons* un système qui *soit* plus clair que le système actuel et qui montre plus clairement les montants alloués à l'UE.

Figure 7.3: Translations produced by the baseline and the tense-aware systems along with source and reference texts.

7.4 Conclusion

In this chapter, we have proposed a fully automatic method for high precision alignment of verb phrases. Even though the method selects only about half of the verb phrases in a text, the large number of occurrences that is available still ensures a large resource. Manual evaluation of a sample showed that about 90% of the labeled occurrences receive a correct label. Similarly to the translation distribution of pronouns, we have found that English verbal tenses present several translation possibilities in French. Concerning the English simple past in particular, we have confirmed the translation ambiguity of this tense as *passé composé*, *passé simple*, and *imparfait*, reported before, and also as *présent*.

In addition, we have evaluated the usefulness of the tense labels to disambiguate the translation of the English simple past into French. We have compared two systems, one trained using the French labels as factors of the English simple past verbs and a second one train without any labels. The system with the labels improved the quality of verbal translations in 0.5 BLEU points, representing ca 27% of the sentences with this particular tense.

Chapter 8

Tense translation modeling: exploiting the bounded/unbounded distinction

8.1 Introduction

In this chapter we investigate the usefulness of the bounded/unbounded distinction to improve the translation of the English Simple Past (SP) into French using a phrase-based statistical machine translation system (SMT). The bounded/unbounded distinction is also called actualization aspect. This aspect is concerned with a distinction between two possible ways of representing or interpreting a particular instance or actualization of an event in a sentence according to its temporal boundaries. It therefore concerns the whole clause or sentence in which a verb is used (Declerck 2007).

As established in the preceding chapter, the four most frequently used translations of the SP in French are: *passé composé* (PC), *imparfait* (IMP), *passé simple* (PS) and *présent* (PRES). We showed that this mapping has a skewed distribution in favour of the PC translation. Since SMT systems favour the most frequent translation and make little use of context or meaning, the other three possible translations are at risk of being undergenerated. This yields translations that can be in some cases ungrammatical and in other cases grammatical but not native-like. As an illustration, in Figure 8.1 we show a sentence translated into French by a baseline system built for our experiments (Section 8.5). We also show a translation by Google Translate. [1] In the example, the italicized

[1] https://translate.google.com/#en/fr/. Translated on August 18th, 2016.

verb is an English SP verb translated using the French PC by both systems. However, the reference translation proposes a PRES form.

SOURCE	It *was* not uncommon for cattle-rustling to occur between cattle-keeping tribes.
PHRASE-BASED SMT	Il *a été* une pratique courante pour vol de bétail lorsque l'on a affaire entre cattle-keeping tribus.
GOOGLE TRANSLATE	Il n'*a* pas *été* rare pour vol de bétail de se produire entre l'élevage du bétail tribus.
REFERENCE	Les vols de bétail ne *sont* pas rares entre tribus d'éleveurs.

Figure 8.1: Example outputs of SMT systems.

Our hypothesis is that the bounded/unbounded distinction is relevant to disambiguate French translations of the English SP. With this assumption, the main goal of this chapter is to integrate this temporal property into a SMT system in order to improve the translation of the English SP into French. To achieve this, we first train a classifier on a manually annotated corpus for predicting *bounded/ unbounded* labels. We the use the classifier to annotate a large corpus with *bounded/unbounded* labels. Finally, we build a SMT system using the automatically annotated corpus in a similar manner to the experiment presented in the previous chapter.

8.2 Aspect and tense

The meaning of a verb or an event is an abstract concept or type. It consists in some abstract representation that individual speakers may share. The notion of aspect, in general, concerns the different ways of viewing an event (Comrie 1976; Declerck 2007), and the actualization aspect, in particular, accounts for *how* that verb or event is actually presented by the speaker in an utterance or in a sentence (Moens and Steedman 1988; Declerck 2007).

The actualization aspect is concerned with a distinction between two possible ways of representing a particular instance or actualization of an event. An event is represented or actualized throughout the clause or sentence as bounded or unbounded.

Following Depraetere (1995) and her examples below, a sentence is considered bounded if it represents a situation as having reached a temporal boundary, irrespective of whether the situation has an intended or inherent endpoint or not. It is unbounded if it does not

represent a situation as having reached a temporal boundary. The examples in (1a) to (1d) contain bounded sentences, those in (1e) and (1f) contain unbounded sentences.

(1) a. I met John at 5 o'clock.
b. Judith played in the garden for an hour.
c. Julian lived in Paris from 1979 until May 1980.
d. I have lived in Paris.
e. She lives on the corner of Russel Square.
f. She is writing a nursery rhyme.

We underline that this type of aspect is a property of the sentence, of the context in which the event is used. The sense or meaning of a verb in isolation is usually compatible with several semantic interpretations[2] (Moens and Steedman 1988, p. 17). In this sense, contextual information from the sentence constrain how an event is to be interpreted. For instance, the modifier in the noun phrase in the sentence '*I ate several apples*' encourages a bounded interpretation, while '*I ate apples*' yields the unbounded interpretation. The first is argued to have a clear endpoint, but not the second. Directional prepositional phrases also affect an event interpretation: '*John pushed the car*' is considered unbounded, while '*John pushed the car into the barn* ' is considered bounded. (Depraetere 1995, pp. 9-11). In fact, one of the linguistic tests that may be used for judging the context of an event as bounded or unbounded is the compatibility with *in* or *for* temporal adverbials (which are prepositional phrases as well). The first indicates a culminated process (2), the second indicates *unbounded* processes like (3).

(2) Laura reached the top in two hours.

(3) John worked in the garden for five hours.

Grammatical aspect, on the other hand, is defined morphologically, in the sense that it refers to particular verbal forms, it refers to the pairing of a forms with meanings (just like grammatical tense) (Declerck 2007, pp. 52-53). Not all languages have a marker for all the aspectual meanings in this sense (Declerck 2007). The only grammatical aspect

[2]The semantic interpretations refer to Vendler (1957)' classification of events as *states* ('love', 'know'), *activities* ('run', 'push a cart)', *accomplishments* ('run a mile', 'draw a circle') and *achievements*. This classification is also known as aktionsart. The definition of these categories along with their interactions with the bounded/unbounded distinction and tense is beyond the scope of our work.

marked in English, for example, is the perfective and imperfective meaning with the -ing suffix.

The choice of a particular verbal tense depends on the fine-grained temporal interpretations of the event in context. The specific placement of the event with respect to its reference time, along with the viewpoint of the situation will determine the particular verbal tense that will be used in the final sentence.

8.3 Data

The corpus used in the experiments that follow was provided by Grisot and Cartoni (2012). The corpus consists of parallel English-French data. The texts that compose it were randomly selected and belong to the following genres: literature, journalism, discussions of the European Union Parliament (the Europarl corpus) and European Union legislation (the JRC-Acquis Corpus). It is a corpus of 435 sentences containing at least one SP token and manually annotated with the following linguistic information: grammatical aspect, narrativity, boundedness, verbal tense used in French and the infinitive form of the verb. Detailed information on the annotation procedure of this information can be found in Grisot (2015).

Meyer, Grisot, and Popescu-Belis (2013) previously made use of this corpus and the *narrativity* information in their MT experiments. Here, we use the same corpus and focus on *boundedness* and its utility for determining the verbal tense used in French as target language. In total, the data contains 236 SP tokens annotated as *bounded* and 199 annotated as *unbounded*, that is 54% and 46% respectively. Most frequently, bounded events correspond to a translation with a PC or PS and unbounded events correspond to a translation with an IMP for 360 items (82%), as illustrated by the first two examples in Figure 8.2. The less frequent cases, namely bounded events corresponding to a translation with an IMP and unbounded events corresponding to a translation with a PC or PS are illustrated in the last two examples in Figure 8.2. This lack of perfect one-to-one correspondence points in favour of a non-deterministic MT system and discourages rigid constraints of the type *bounded* → PC, *unbounded* → IMP, such as those used in the system described by Olsen et al. (2001).

With only 435 sentences, the annotated corpus is very small and therefore not sufficient to train a SMT system. For this reason, as described in the next section, the manually

Sentence	Verb	Infinitive Aspect	Grammatical	Narrativity	Bounded-ness	FR tense
In one instance, Kazakhstan revealed the existence of a ton of highly enriched uranium and *asked* the United States to remove it, lest it fall into the wrong hands.	asked	to ask	perfective	narrative	bounded	PC
He *fascinated* everybody who was worth fascinating and a great number of people who were not. He was often wilful and petulant, and I used to think him dreadfully insincere.	fascinated	to fascinate	imperfective	non-narrative	unbounded	IMP
A few days ago, in a manner of speaking, we *said* that Bin Laden had provided the impetus for implementing methods for fighting terrorism that the Commission had been planning and that Parliament had requested some time ago.	said	to say	perfective	narrative	bounded	IMP
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.	were seen	to see	imperfective	non-narrative	unbounded	PC

Figure 8.2: Example of the annotated corpus data provided by Grisot and Cartoni (2012).

annotated corpus is used as training data to build a classifier which annotates new data with *bounded/unbounded* labels. The following sections are dedicated to describing, on the one hand, experiments on the prediction of *bounded* and *unbounded* labels for English SP verbs, and on the other hand, experiments with a SMT system trained on a corpus annotated with *bounded/unbounded* labels.

8.4 Predicting the *boundedness* of English SP verbs

We chose to work with the *bounded/unbounded* distinction for two main reasons: availability of resources and the need for a disambiguation factor of the English SP tense into French which could be predicted from the information in the sentence and not the verb itself, considering that SP tokens do not have disambiguation morphology themselves. In the experiments presented below, we assume that the information available in the sentence where the verb occurs can be used to predict the boundedness status of English SP verbs due to its context-dependent character.

8.4.1 Experiment 1: Using all relevant features

In this first experiment, a classifier is trained for predicting the type of *boundedness* of the English SP verbs from the manually annotated corpus. The additional linguistic annotations of the corpus are exploited as features for the classifier, based on Grisot (2015) who proposes that they are pertinent for the task. Additional syntactic and temporal features automatically generated and extracted from the sentence in which the verb occurs are also included. Since this classifier is partially fed with features known to be pertinent for the task, its results are expected to be a measure of the maximum success rate on this particular task and using this data.

The Stanford Maximum Entropy package is used to build the classifier. The entire manually annotated corpus is used both as training and testing data. Given its small size, results are reported as averages over ten-fold cross-validation for the two experiments which follow. Note that the ten-fold validation ensures low variance and maximum generalization power given that the corpus is very small. *Boundedness* is the prediction class and it has two possible values *bounded* or *unbounded*.

Features

The features used are of two types: syntactic and temporal. Syntactic features model the context (i.e. the sentence) in which the English SP verb occurs, whereas temporal features refer to the interpretation or meaning of the SP verb itself. Manually annotated features which were taken from the previous corpus annotation scheme are indicated with an asterisk (*) symbol. For the automatically generated features, the dependency parser from MateTools was used on the English side of the corpus to produce POS-tags and dependencies labels.

Syntactic features

1. Position in the sentence: refers to the ordinal position of the English SP verb in the sentence.
2. POS-tags of the English SP token: they distinguish between active voice SP verbs, e.g., *went* (VBD); compound active voice SP verbs e.g., *did go* (VBD+VB); and passive voice SP verbs, e.g., *was taken* (VBD+VBN).
3. Head and its type: it refers to the syntactic head of the verb to classify, along with

its POS-tag.

4. Children dependencies: they indicate the dependency relation of the three nearest children of the English SP verb.
5. Children POS-tags: they indicate the POS-tags of the three nearest children of the verb. With this and the previous feature, we expect to capture some of the linguistic reflexes of aspect (Section 8.2), for example the presence of *in* prepositional phrases for *bounded* eventualities.

Temporal features

6. Simple past verb token*: this refers to the English SP verb to be classified.
7. Infinitive form*: the non-finite form of the English SP verb token.
8. Grammatical aspect*: a pragmatic feature taking the values of *perfective*, which stresses the initial and final boundaries of an eventuality, or *imperfective*, which does not stress these boundaries.
9. French tense*: the tense of the French translation corresponding to the English SP verb in the parallel corpus.
10. Adverbs: Meyer, Grisot, and Popescu-Belis (2013) manually gathered a list of 66 adverbial (temporal) expressions; we checked for the presence or not of such expressions in the English sentence.
11. The type of adverb: additionally, each adverbial expression was labeled by Meyer, Grisot, and Popescu-Belis (2013) as a marker of synchrony (e.g., *meanwhile*) or asynchrony (e.g., *until*). These type labels were also included among the features.
12. Narrativity*: a pragmatic feature referring to the temporal structure between eventualities. It can have the values of *narrative* or *non-narrative*.

Results

Results show a very good performance of the classifier, reaching up to 0.8943 F-score for the bounded class and 0.8759 for the unbounded class. These results are partially explained by the features taken from the previous annotation of the corpus, produced by expert linguists. However, even if all features are pertinent and linguistically-motivated, they are not error-free. Those generated using an automatic tool in particular may intro-

	Bounded	**Unbounded**
Precision	0.8833	0.8909
Recall	0.9038	0.8650
F1	0.8943	0.8759
Accuracy	0.8857	

Table 8.1: Average classification results of Experiment 1 using ten-fold cross-validation.

duce some amount of noise. In what concerns the gold annotation of the *bounded* and *unbounded* labels, they contain some degree of ambiguity as well. As expressed by annotators, judgments can be ambiguous since they also depend on the particular context each verb appears in. These results reflect to some extent the intrinsic ambiguity of the *boundedness* of English SP verbs.

8.4.2 Experiment 2: Using limited features

The main goal of our work in this chapter is to enhance a SMT system with *boundedness* information as a means to disambiguate English SP verbs when translating into French IMP, PRES, PC or PS. For building a SMT system, a 435 sentences corpus is clearly insufficient, a much larger parallel corpus is needed. As in the previous experiment, here a classifier is trained for predicting one of the two values for *boundedness* of the English SP verbs. However, the objective of this second classifier is to approximate the results obtained in Experiment 1, using a sub-set of the features previously described in 1 to 12. This sub-set is composed of those features which it is possible to generate from raw data. Consequently, the results of this experiment are expected to give a realistic impression of the quality of the *boundedness* detection task on a large corpus using automatically generated features and a small quantity of annotated data –the only annotation being the gold prediction class– for training.

As in Experiment 1, a Maximum Entropy classifier is built using the Stanford Maximum Entropy package. The dependency parser from MateTools is used on the English side of the corpus to obtain dependency relation information. Additionally, in this experiment, the TreeTagger system which produces POS-tags and lemmas for all words in the sentence is used on the English side of the corpus as well. The complete manually annotated corpus is used as training and as testing data. Results are reported as averages over ten-fold cross-validation. As before, *boundedness* is the prediction class and it has

two possible values *bounded* or *unbounded*.

Features

Previously, the manual annotations already existing in the corpus was recovered as features for the classifier since they were known to be pertinent for the task. Some of those features can be easily obtained using syntactic and morphological parsers. However, this is not the case for *grammatical aspect*, *French tense* and *narrativity*. In this second Experiment, the input to the classifier is limited to the features which will be available when using the parallel SMT data, those created automatically. Following the same intuition as before, the training features are divided into syntactic and temporal types.

Syntactic features

1. Position in the sentence
2. POS-tags of English SP token
3. Head and its type
4. Children dependencies
5. Children POS-tags

Temporal features

6. Simple past verb token: this refers to the English SP verb to be classified. In this experiment, we used the heuristics based on POS-tags described in the previous chapter to identify all English SP instances in the sentence.
7. Infinitive form: the non-finite form of the English SP verb. It was extracted from the output produced by TreeTagger.
8. Temporal adverbs
9. The type of adverb

Results

Table 8.2 shows the results. Note that in the first experiment, as in the corpus, one SP verb per sentence is annotated. In this experiment we identified all English SP instances in a sentence automatically. Nonetheless, results reported in Table 8.2 are limited to

	Bounded	**Unbounded**
Precision	0.8142	0.8509
Recall	0.8747	0.7578
F1	0.8401	0.7944
Accuracy	0.8224	

Table 8.2: Average classification results of Experiment 2 using ten-fold cross-validation.

the same verbs annotated originally and used in Experiment 1. Hence, results of both experiments are comparable. For the subsequent SMT experiments, all English SP verbs are identified and tagged as *bounded* or *unbounded*.

The quality of the classifier is quite satisfactory, reaching up to 0.8401 F-score for the bounded class. Results are comparable to those of Experiment 1. In this experiment, however, the unbounded class seems harder to predict than in Experiment 1, as evidenced by the generally lower figures, recall in particular.

8.4.3 Discussion

Experiment 1 showed that *boundedness* could be predicted from sentence features. These features were partially annotated by hand and they were expected to be relevant for the task. Experiment 2 produced good quality results despite the partially missing gold information used in Experiment 1 (i.e., grammatical aspect, French tense, narrativity). While Experiment 1 set the upper bound of the task, the results of Experiment 2 were established under more realistic conditions, since automatic tools were used to generate the features (which also implies some noise). The second experiment measured the quality with which completely raw data can be automatically annotated. There is a significant difference of about 8% in performance between the two classifiers ($\tau(434) = 7.28$, p-value = 1.5e-12). Yet, the second classifier was still able to learn how to discriminate between *bounded* and *unbounded* SP verbs.

To measure the impact of the result, we set a baseline based on randomisation for comparison. A random sample with resampling of 435 *bounded/unbounded* labels with probabilities 0.54 and 0.46 respectively was generated. These probabilities correspond to the distribution of the labels in the human annotated corpus (Section 8.3). Next, we compared the obtained random labels to the gold corpus in order to compute precision, recall and F-score in the standard fashion (Table 8.3). Both the results of Experiment

1 and Experiment 2 are significantly better than our random sample (τ(434) = -76.71, p-value = 2.2e-16; τ(434) = -57.05, p-value = 2.2e-16), which further indicates that the prediction results are solid. A graphical summary of this comparison is given in Figure 8.3.

	Bounded	**Unbounded**
Precision	0.5574	0.5192
Recall	0.5763	0.4426
F1	0.5667	0.4779
Accuracy	0.5402	

Table 8.3: Results of a sample of 435 randomly generated labels according to their gold distribution probability.

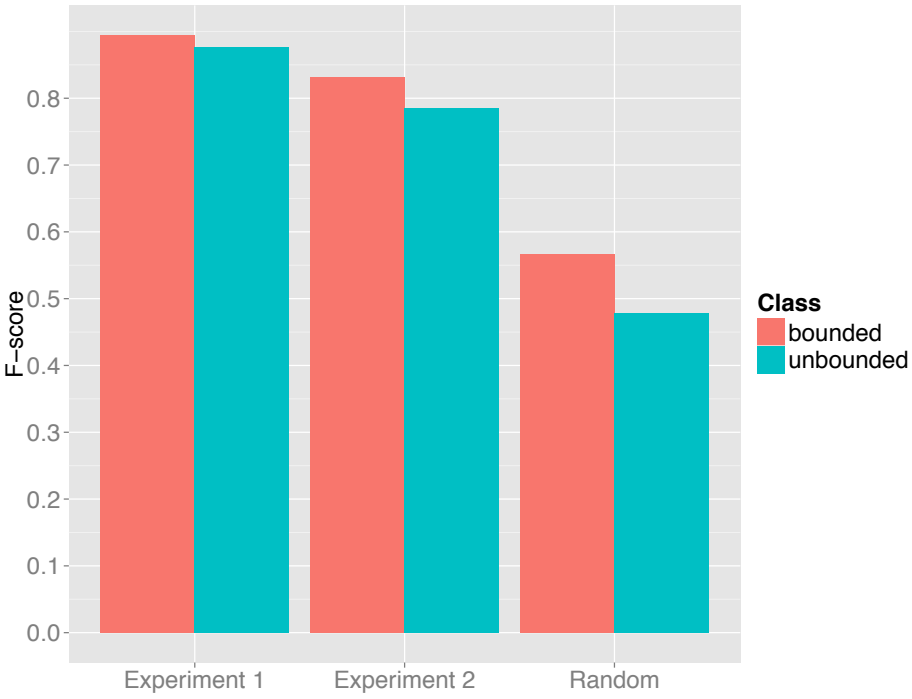


Figure 8.3: Comparison of results obtained in the classification Experiments. The blue represents the *bounded* class and the red represents the *unbounded* class.

To judge the predictive power of each of the features involved, feature ablation for each of the experiments was done. We compared the performance of the classifier trained on all the features to its performance when each feature is subtracted (one at the time) from the model. For each feature removal round, we used ten-fold cross validation and

calculated the F-score for each class. The observed changes are plotted in Figures 8.4 and 8.5; the mean of all the folds is given by the thick middle line in each boxplot.

In both experiments, the interaction of the features is dependent on the class to be predicted. For example, grammatical aspect and narrativity seem to be important for the unbounded class only, while the verb's POS tags seem to be more informative for the bounded class. However, it is clear that the adverbs and the infinitives are the features with the most predictive power for both classes and in both experiments. Moreover, although we initially thought that the verb position could be an indicator of main (lower values) vs subordinated verb status (higher values), the analysis of the results indicated that it is not very informative. The verb's children dependencies are another feature which did not provide improvements to the model. The children POS tags is a more useful feature.

Following Meyer, Grisot, and Popescu-Belis (2013), we also experimented adding information concerning the previous verb to the feature vector of the current verb. The intuition was that modeling more context would benefit the classifier. We noted, however, a decrease in accuracy of around 15%, therefore this information was dropped. This outcome is in line with what is reported by Ye, Schneider, and Abney (2007, p. 6) on aspect prediction for Chinese: "We expanded the features vector of each verb by including the features from the previous and the following verb; the results showed no improvement by adding these contextual features." In our case at least, this outcome might be due to data scarcity, given the small size of the corpus.

To conclude, the results and analysis presented in this section indicate that the classifier from Experiment 2 is satisfactory and in the next section we will test if its quality is sufficient to improve SMT. Since this classifier is trained exclusively on features created automatically, it serves our purpose of annotating training data to build a SMT system enhanced with aspect information. In the next section, the classifier from Experiment 2 is used to annotate all English SP verbs in a large corpus with *bounded* and *unbounded* labels in order to train a SMT system. The purpose of the SMT experiment is to test the disambiguation capability of the boundedness property for the translation of the English SP. Boundedness labels should be used by the SMT to improve the choice of the verbal tenses in French.

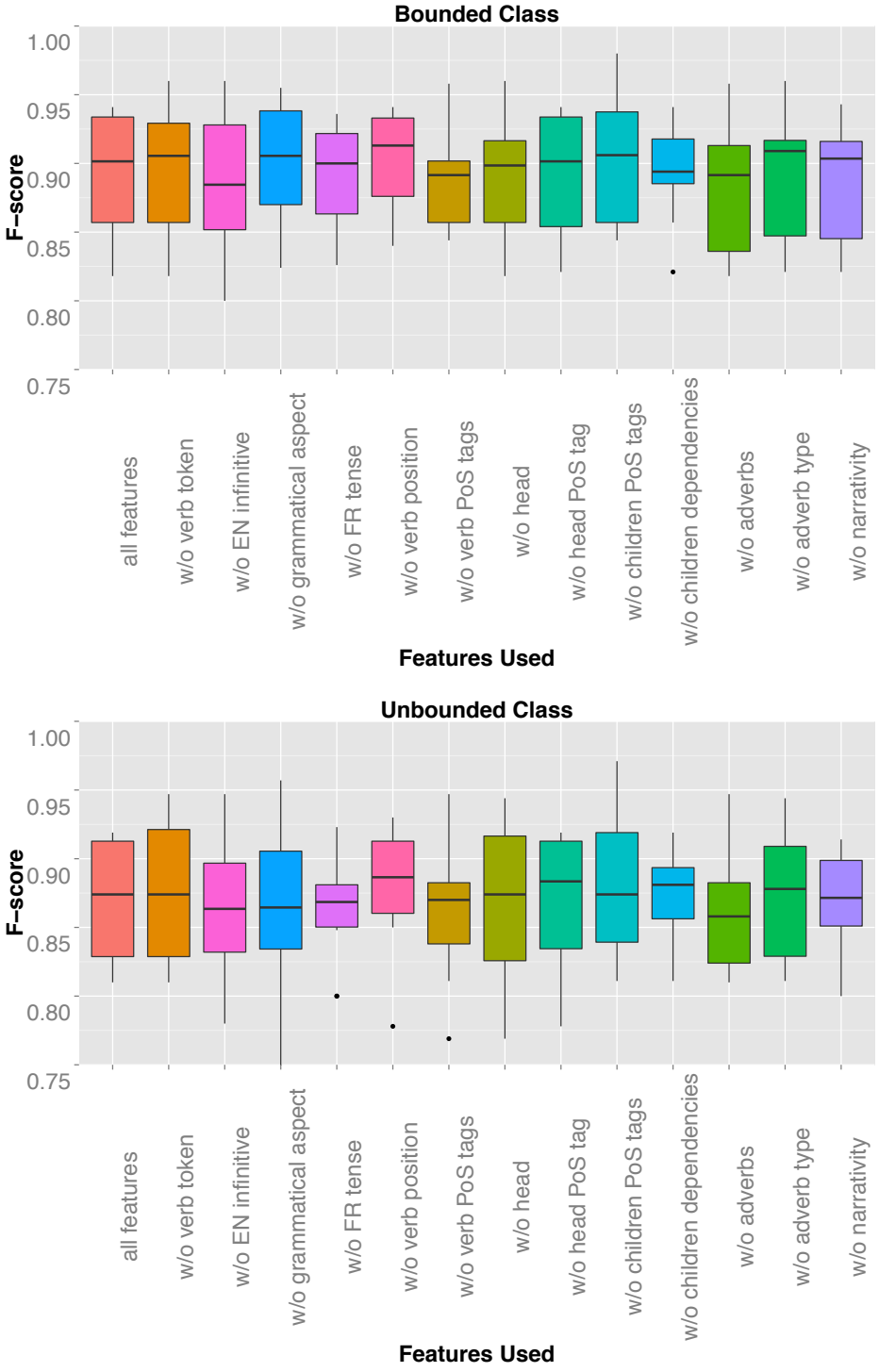


Figure 8.4: Feature ablation comparison for Experiment 1.

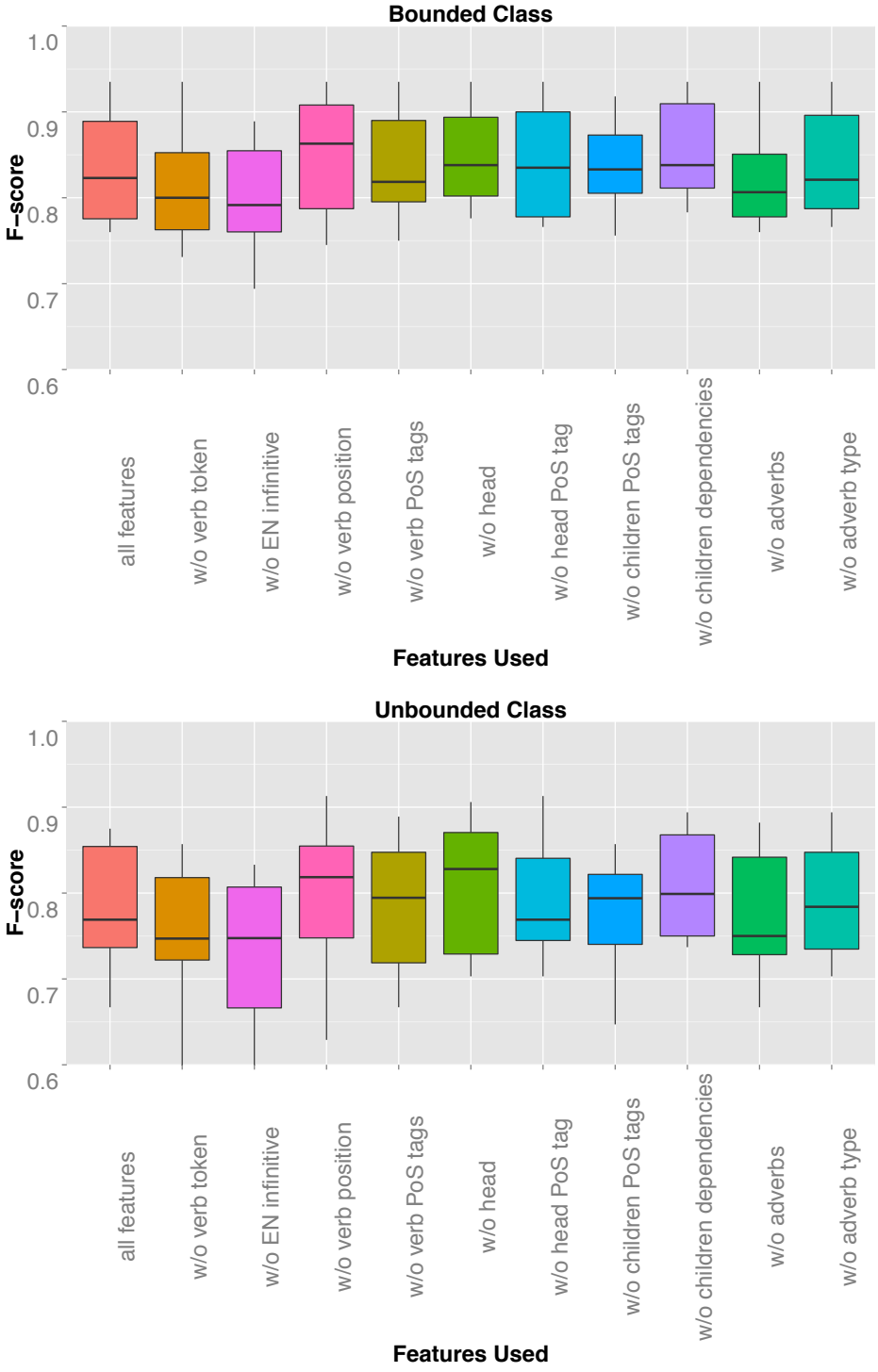


Figure 8.5: Feature ablation comparison for Experiment 2.

8.5 Using the predictions for machine translation

In this final part, we assess a SMT system enhanced with *bounded/unbounded* labels and the translation quality of the English SP into French. The PC translation possibility is the most frequent one, increasing therefore the chances of being generated a 'default' translation by phrase-based systems. As in the previous chapter, our goal is improve the verbal tense translation choices of the system by boosting the other translation possibilities, PS, IMP and PRES.

We annotate the training data using the classifier built before in Experiment 2 (Section 8.4). The data is taken from the MultiUN corpus, a corpus of translated documents from the United Nations, provided by Eisele and Y. Chen (2010). All English SP verbs are identified and labeled as either *bounded* or *unbounded* automatically. Table 8.4 shows the number of English SP verbs annotated with this method.

	Sentences	**SP verbs**
Training	350,000	134,421
Tuning	3,000	1,058
Testing	2,500	1,275

Table 8.4: Data setup of the SMT system.

We use the Moses Toolkit to build two systems: a baseline without *bounded/unbounded* labels and an aspect-aware system with the labels. Both systems are phrase-based models with an identical composition, according to the set-up presented in Table 8.4, and use a 3-gram language model built using KenLM (Heafield 2011). The language model is trained over 10 million sentences of French monolingual data taken from the 2015 Workshop on Machine Translation (WMT15) (Bojar, Chatterjee, et al. 2015). Optimization weights were fixed using Minimum Error Rate Training (MERT) (Och 2003). We use a relatively low number of training sentences in order to facilitate comparison with our work in the previous chapter and with previous research by Meyer, Grisot, and Popescu-Belis (2013).

The *boundedness* labels are combined with the SMT system using a factored model (Koehn and Hoang 2007). In our system, only one factor, i.e. the bounded or unbounded label is used. All the other words have NULL as default factor. For verbs composed of multiple words, (e.g., *cut off, was made, were laid down*) all words are labeled with the same *bounded* or *unbounded* factor.

In the model, factors are taken into account in a non-deterministic manner. In other

words, there is no exact mapping between a given label and a particular output. For instance, a *bounded* label does not necessarily lead to a verb with a PC French tense. Instead, factors are considered when estimating the translation probabilities computed over the entire parallel corpus.

8.5.1 Results and discussion

The results obtained are given in Table 8.5 using the BLEU (Papineni et al. 2002) score. The system with the *boundedness* labels (aspect-aware) obtained an increase of 0.98 BLEU points. When computing the BLEU score on the sentences with SP verbs only, we obtained a difference of 1.58 points. These scores reflect an improvement in the quality of the SP translation. On the one hand, these increments suggest that the method is not degrading the general translation quality of all the other words in the sentence; on the other hand, they suggest that it is not changing the SP translations estimated as already adequate by the baseline model. This result is non negligible, considering in particular that the aspect-aware system targets only SP verbs and not all words in the sentence. BLEU, being an exact-matching-oriented metric, increases as the number of matching words to the reference increases.

System	BLEU test set	BLEU SP sub-set
Baseline	31.75	30.05
Aspect-aware	32.73	31.63

Table 8.5: BLEU scores of the SMT systems computed on the test set and and on the sentences with SP verbs only.

As with the systems in the preceding chapter, a bootstrap with resampling significance test as introduced by Koehn (2004) was carried out. We took 100 paired samples of 300 sentences from the test set, each containing at least one verb in the SP tense. Then a BLEU score is computed and recorded for each sentence in each sample. Results are given in Figure 8.6. Consistently across all the samples, $\approx$ 50% of the sentences containing at least one English SP verb were better translated by the aspect-aware system than by the baseline system. Furthermore, following the method proposed by Zhang, Vogel, and Waibel (2004), a 80% confidence level estimate computed over the 100 samples places the confidence interval of the differences in BLEU scores between the two systems at [0.86, 2.68]. In other words, we are confident that 80% of the samples of a

system trained with the labels will be at least 0.86 BLEU points better than a system without the labels.

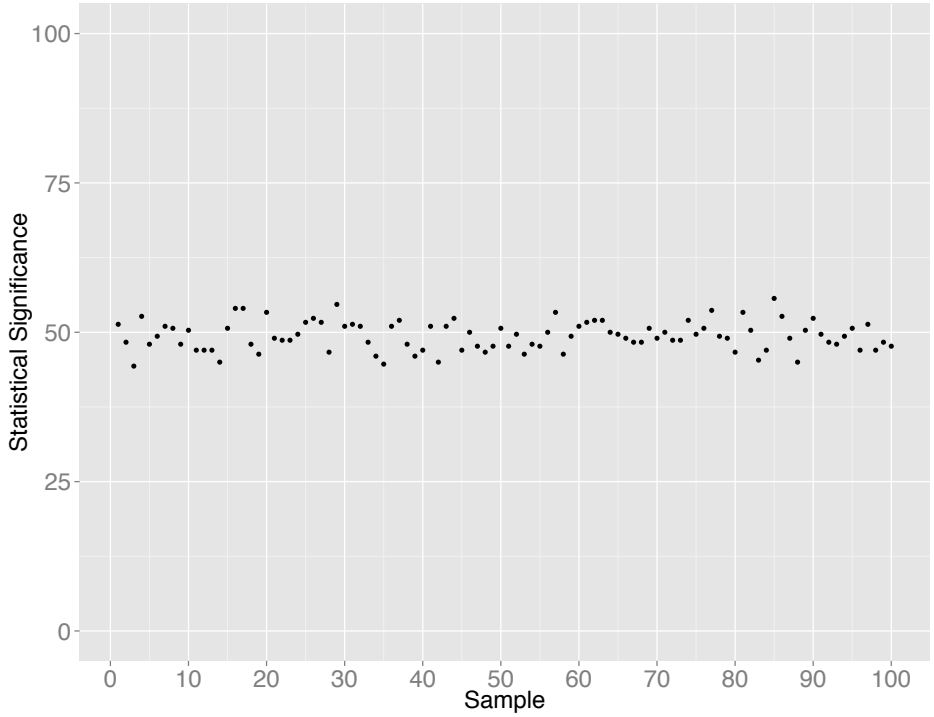


Figure 8.6: Results of paired bootstrap with resampling test. The x axis shows the ID of the samples and the y axis displays the percentage of sentences per sample which obtained a higher BLEU score than the sentences in the baseline.

To gain more insights on the qualitative differences in the tense translation between the outputs, we completed a manual evaluation of 200 randomly selected English SP verbs as well. The selection contains an even distribution of the labels. Results are summarized in Tables 8.6 and 8.7.

Table 8.6 shows the assessment of the classifier performance. The verb boundary identification is very good with 91% accuracy. As mentioned, verb identification was done automatically, using POS-tags along with a set of heuristics from Chapter 7. Errors are mostly due to incorrect tagging of some ambiguous cases such as the construction *was concerned* in which only *was* is identified. Another common error occurred with adjectives identified as verbs, for instance *titled* in '*... under the item titled "the Situation in the Middle East"*'. For the labeling, the manual evaluation shows 65% accuracy, a figure lower than the results obtained automatically and presented in Table 8.2.

	Correct	Incorrect	Total
Verb Identification	182 (91%)	18 (9%)	200
Predicted Label	117 (65%)	65 (35%)	182

Table 8.6: Manual evaluation of the classification results of 200 SP verbs. The predicted label is evaluated only on those cases where the SP is identified correctly.

In general, the bounded class seems more difficult to predict than the unbounded one. The manual evaluation also revealed that several verbs which usually express one-time events, as *ask*, *request*, *result*, *adopt*, *add* or *call*, were treated as having a duration which is much less common. Finally, we noticed that several instances of the same verb appear repeatedly, therefore, the same classification error was repeated (e.g. *was* labeled as bounded).

Table 8.7 presents the results of the comparison between the baseline and the aspect-aware systems. The classification of the English SP verbs was correct in 65% of the cases and their translation is improved in 25% of the cases. Most verb translations are unchanged, most probably because the weight of the *bounded/unbounded* factor yields the same best translation hypothesis as the baseline, in other words, the same translation probability would be produced without the factor. Indeed, the translation distribution is highly skewed in favor of the PC. Last, 21% of the examples were degraded, a possible outcome given the non-deterministic disposition of the factored model. This result is also directly linked to the results of the *bounded/bounded* labeling: correct labels entail twice as many improved translations (31 vs 17).

Bounded/ Unbounded	**Translation**		
	Improved	**Unchanged**	**Worsened**
Correct	31	69	17
Incorrect	15	29	21
Total	46 (25%)	98 (54%)	38 (21%)

Table 8.7: Relationship between classifier results and translation quality of the 182 correctly identified SP verbs.

Two examples of the improved translations are presented in Figures 8.7 and 8.8. In Figure 8.7, the verb *was* is labeled as *unbounded* and translated in French using the IMP by the aspect-aware system. This is the same translation used in the reference. The baseline system produces a PRES tense translation. In Figure 8.8, both verbs *had* and *were* are labeled as *unbounded*. In this example, however, the labeling seem to have

an effect only on the first one. Both verbs are translated using the PC by the baseline. The aspect-aware model, instead, produces the IMP tense (same as the reference) for the first one but not for the second. Other differences between the translation outputs in this example are probably due to a different ranking of the hypothesis during decoding time. Different hypothesis combination is likely to happen when translating long sentences.

SOURCE	Education *was* mandatory up to the age of 16.
BASELINE	L'éducation *est* obligatoire jusqu'à l'âge de 16 ans.
ASPECT-AWARE	L'éducation *était* obligatoire jusqu'à l'âge de 16.
REFERENCE	L'enseignement *était* obligatoire jusqu'à l'âge de 16 ans.

Figure 8.7: Example outputs of the SMT systems.

SOURCE	He also considers that he has exhausted domestic remedies with regard to release on bail, and that the remedies mentioned by the State party *had* no prospect of success and *were* not available.
BASELINE	Il considère aussi qu'il a épuisé tous les recours au niveau national en ce qui concerne libération sous caution et que des procédures de recours mentionnée par l'État partie *n'ont* pas de chances d'aboutir, et n'*ont* pas *été* communiquées.
ASPECT-AWARE	Il estime qu'il a épuisé les recours internes en ce qui concerne la libération provisoire sous caution, et que les recours mentionné par l'État partie n'*avaient* aucune perspective de succès et n'*ont* pas *été* disponible.
REFERENCE	Il estime également avoir épuisé les recours internes quant aux demandes de mise en liberté sous caution, et que les recours mentionnés par l'Etat partie n'*avaient* aucune chance d'aboutir et n'*étaient* pas disponibles.

Figure 8.8: Example outputs of the SMT systems.

8.5.2 Comparison with previous research on narrativity

The method presented in this chapter has been used by Meyer, Grisot, and Popescu-Belis (2013) and Loáiciga, Meyer, and Popescu-Belis (2014) on the same problem to test different types of temporal properties. As mentioned earlier, Meyer, Grisot, and Popescu-Belis (2013) previously made use of the corpus provided by Grisot and Cartoni (2012) and the *narrativity* annotation. In Loáiciga, Meyer, and Popescu-Belis (2014), we used the tense annotation presented in the previous chapter to train a classifier to

Temporal information	Δ confidence interval	Δ in BLEU
Narrativity	n/a	+0.2
Tense-predicted	n/a	+0.12
Tense-oracle (this work)	[0.1, 0.8]	+0.5
Boundedness (this work)	[0.8, 2.7]	+0.9

Table 8.8: Comparison of reported gains in the works on disambiguation of the translation of the English SP into French.

annotate the data for a SMT system with tense labels. In the present work, we use Grisot and Cartoni (2012)'s corpus and focus on *boundedness*. We summarize the reported gains with respect to a baseline without the pertinent temporal information in Table 8.8.

Narrativity is a pragmatic property which refers to the the temporal relations holding among events. Two cases are possible: *narrative* and *non-narrative* usages. A narrative usage points to the case when the two events are temporally linked (with both forward and backward temporal inferences). Non-narratives usages point to the case when events are either not temporally linked or they occur simultaneously. *Grammatical tense* is a morphological feature expressed in the pairing of different temporal meanings with different verbal forms. Last, *boundedness* refers to an aspectual property of the event used in context, it refers to a property of the sentence in which the verb occurs. It can be seen from the comparison of results above, that it is this last type of temporal information which produces the biggest gains in terms of the disambiguation of the French translation of the English SP. Since this is a property of the context in which the verbs appears, it could be applied to other language pairs such as Chinese-English, in which there is great temporal ambiguity as reported by Ye, Fossum, and Abney (2006) and Ye, Schneider, and Abney (2007).

8.6 Conclusion

In this and the previous chapter, we have experimented with the constructs of tense and aspect to disambiguate and improve the automatic translation of the English SP verbs into French. This single English tense has four major possible translations in French.

We have built two classifiers for *boundedness*, one including a larger set of features including oracle features and the other one trained on automatic features only. The first showed that boundedness can be annotated reliably and set the upper-bound performance

of the classification task. The second allowed us to label a large corpus based on a minimal and affordable quantity of manually annotated data. Regarding the classification tasks, we found that training on such a small corpus produced good results. Compared to other latent features difficult for automatic prediction such as narrativity or aspect markers in Chinese, the component of aspect that we examined seems more feasible to learn. We obtained results around 15% better than those for narrativity prediction (Meyer, Grisot, and Popescu-Belis 2013) for instance.

The method proved to have good results with respect to the targeted verbal tenses without decreasing the quality of the translation of the surrounding words in the sentence. Indeed, manual evaluation of the translated texts showed that correctly labeled verbs with boundedness presented a better tense translation. With this work, we hope to have contributed to building a more natural and cohesive MT output.

Chapter 9

Conclusions and future directions

In this chapter, we take a step back and recapitulate on what has been achieved in this work before reflecting on future avenues of research. This thesis has presented experiments aimed at including linguistic knowledge in ways useful for creating better machine translation output within existing machine translation architectures. Two referential devices, pronouns and verb tenses, have been targeted for their contribution to textual cohesion and coherence.

Concerning pronouns, we have investigated whether syntactic knowledge is relevant for pronoun translation in the form of binding rules in a classic anaphora resolution framework, and in the form of features for cross-lingual pronoun translation. Concerning verb tenses, we have investigated whether the constructs of grammatical tense and actualization aspect had a positive effect on translation quality.

Rule-based, statistical and deep learning methods have been used throughout the experiments, reflecting the complexity involved in the translation of both, pronominal and verbal reference.

9.1 Conclusions

Pronouns and verb tenses, the two linguistic phenomena treated in this thesis, facilitate textual reference. This means that they both function in association with another element, an antecedent or referent, to express meaning, and that either a pronoun-antecedent pair or a tense-referent pair are potentially placed in different sentences. Such character-

istics cause that their translation must take into account the lexical, morpho-syntactical and discoursive linguistic levels. For pronouns, the lexical properties of the languages involved, as well as their principles of morphological agreement and syntactic binding interact. For verb tenses, the interpretation cues for their correct understanding come from several elements as diverse as the lexical choice, the adverbs and the particular type of clauses used.

Concerning the translation of pronouns, ideally, the machine translation process should preserve the sense intended in the source language and produce translations which are consistent with the properties of the target language. These properties include not only the grammatical agreement between the pronoun and the antecedent, but also the preferences in usage of pronouns and nominal reference in general. While a pronoun-antecedent pair is not necessarily always the most fluent or the most adequate translation, other translations should be considered, whenever a pronoun-antecedent pair is used, the agreement features must be grammatical.

We have shown that the distribution of pronouns in corpora depends on the language and the genre of the text. In the case of the translation between English and French, we have also extensively demonstrated that pronouns are particularly susceptible to translation variations, and that one-to-one translations or even pronoun-to-pronoun translations are by far not the only translation possibilities.

We looked into two different approaches for pronoun translation from English into French: rule-based machine translation with classic anaphora resolution and cross-lingual pronoun prediction without anaphora resolution. Using the Its-2 machine translation system (Wehrli and Nerima 2009) and syntactic theory, we have found that in the former approach, the problem of pronoun translation goes beyond the anaphora resolution problem. Only a subset of the pronouns (*il*, *ils*, *elle*, *elles*) is generated correctly based on the agreement features of their nominal antecedents. Therefore, this solution resulted limited.

The second approach defines a fixed number of classes which includes a OTHER class to account for all the non-pronominal translations. Most of the previous research on cross-lingual pronoun prediction includes different sources of information, in the form of features, with no systematic link between the features and the interaction they may have with each individual class to predict. Working with the Stanford Maximum Entropy package (Manning and Klein 2003), this approach has allowed us to rank the im-

portance of contextual, morphological and syntactical sources of information, for the correct translation of pronouns. Concretely, we have found that the immediate context is the best predictor overall, with a particular good influence over the high frequency classes *il, ce, ils,* and OTHER. Whereas syntactical features are beneficial for the *ce, cela, elles, il* and *ça* classes, morphological features showed a positive effect for the *elle* class only. The *ça* and *cela* classes obtained the worst results overall.

We expected a sharper interdependence between the groups of features tested and the individual classes. We consider, therefore, that further investigation into the types of features used until now for the task (both here and in previous work) would help the definition of a model for English-French pronoun translation.

We have compared our work on pronoun prediction to many systems which use explicit antecedent links from a stand-alone anaphora or coreference resolution system. This type of knowledge, as we confirmed in our own rule-based experiments, helps the translation of just a subset or pronouns, and in many cases, it turns out to be deficient. These systems are only moderately successful at identifying potential antecedents and pairing them with pronouns correctly. Their quality is even more compromised for languages other than English.

In contrast with earlier research, we provide evidence in favour of the usefulness of deep syntactic knowledge to improve the performance of pronoun prediction by an average of 2.45 F-score. We have exploited the syntactic relations between a verb and its arguments (subject, direct, indirect, predicative and sentential objects) in the form of features. Stymne (2016) has followed our example and has also concluded that the syntactic knowledge in the form of dependency relations between a verb and its arguments is beneficial for the task. This finding does not support the previous conclusions by Kehler et al. (2004), in a piece of research which has been pointed out by the coreference resolution community as evidence that linguistic knowledge (syntax in particular) is not relevant for pronoun resolution. Taken together, these results suggest that syntactic knowledge plays a role in pronoun resolution and translation. However, the exact manner in which such information should be utilized in a statistical framework is less clear.

It is worth repeating that contrary to the machine translation of the complete source text, the cross-lingual pronoun prediction approach has the advantage of an easy evaluation. Since a fixed number of classes is defined, the task is evaluated as a standard classification task. While not perfect, the OTHER class accounts for some of the uncertainty of the

translation possibilities.

Our analysis of the role that different types of features play in the prediction of the pronouns studied here has led us to propose a three-way distinction of the function of pronoun *it*: pleonastic, nominal anaphoric and event anaphoric. It is to note that we have built upon work by Guillou (2016) who has presented evidence in favour of incorporating pronoun function into MT and provided the ParCor corpus. Distinguishing between nominal anaphoric and event reference realisations of *it* proved to be a complex task, especially if we consider that, after all, event reference is itself a form of anaphoric reference. For the *it* disambiguation task, the combination of a maximum entropy classifier and a recurrent neural network system resulted in performance gains which reflect their respective strengths. The recurrent neural network system proved better at handling ambiguous referring relationships, while the maximum entropy classifier performed better at identifying clear antecedent-pronoun pairs. So far, our results are encouraging.

Turning now to verbal reference, verb tenses have been argued to be anaphoric and therefore related to the discourse level of language. That a pronoun and its referent hold an intersentential relationship has been shown in many of the examples presented in our work. However, it has not been shown that a verbal form and its referent hold an intersentential relationship. The sequence effect of verb tenses at the text level has been discussed in the related research. However, we have limited our experiments to the sentence level because we worked with a small manually annotated corpus of isolated sentences which makes it impossible to consider the context of the previous sentences.

We have focused on the translation of the English simple past into French. Building on previous small-scale studies, we have presented quantitative evidence in favour of a translation ambiguity of this tense into French as: *passé composé*, *imparfait*, *passé simple* and *présent*. We considered the usefulness of *grammatical tense* and *boundedness* for the translation of the English simple past into French and compare our work with previous research on *narrativity* for the same task (Meyer, Grisot, and Popescu-Belis 2013).

Grammatical tense is a morphological feature expressed in the pairing of different temporal meanings with different verbal forms. Boundedness refers to an aspectual property of the event used in context, it refers to a property of the sentence in which the verb occurs. Narrativity is a pragmatic property which refers to the temporal relations holding among events. A narrative relation points to the case when the two events are tempo-

rally linked an a non-narrative relation point to the case when the events are either not temporally linked or they occur simultaneously. From these properties, our machine translation experiments showed that boundedness produced the best results to improve the translation of the English simple past into French, increasing translation performance up to +0.9 BLEU points. Tense improves the translation performance up to +0.5 BLEU points, whereas narrativity improves it up to +0.2 BLEU points.

Finally, our work has had a strong focus on English-French, mainly because high quality machine translation output is needed when working on targeted issues such as pronouns and verb tenses. English-French is one of the few language pairs for which current machine translation systems obtain high quality. This is supported by the fact that it is no longer included as part of the shared tasks organized by the Conference on Machine Translation (WMT). A system which still produces unintelligible output probably needs more general issues to be tackled first. Low quality translations, in addition, do not allow to detect specific output changes due to the treatment of a subset of the input data. Yet, we stress that the ideas investigated in this thesis can be applied to other language pairs.

First, the rule-based anaphora resolution component can work for all the other language pairs covered by the Fips parser and the Its-2 machine translator. At the present, this system works well for ten language pairs between English, French, German, Italian and Spanish. Besides, the resolution strategy itself could be implemented into any system with a similar architecture.

Second, our pronoun prediction experiments do not include any anaphora or coreference resolution system, which is an advantage as such systems exist mainly for English. Indeed, existing coreference resolution systems seldom work on raw data and corpora annotated with coreference information are rare for any other language than English. From the works reviewed here, only Bawden (2016) reports training a coreference resolution system for French, with rather disappointing results.

Third, the *it* disambiguation task is potentially useful in all translation language pairs where there is a one-to-many third person pronoun correspondence. We have illustrated with a small sample of 58 English instances of *it*, that this pronoun has different translation preferences in French and German, suggesting that this may be the case in other target languages with more than one third person pronoun as well. In addition, we believe that this task could benefit not only machine translation but also the task of coreference resolution. These systems often include different strategies to identify pleonastic

instances of *it* in order not to attempt their resolution. If, additionally, a coreference resolution system identifies event anaphoric instances of *it*, their matching with a specific nominal antecedent could be prevented and their inclusion in a correct coreferential chain of events could be promoted.

Fourth, although the automatic tense annotation method is specific to English and French, a similar annotation approach is possible for other languages where part-of-speech tagging and parsing exists. Although coverage for all the languages of the world is far from existing, parsing models are available for many other languages than English and French.

Last, the bounded/unbounded distinction used to disambiguate the English simple past verbs does not specifically concern English, but the context in which the verb occurs. This means that it could be used for pair of languages such as English-Chinese, where verbal tense translation is a known problem. Language independent properties like the bounded/unbounded distinction are valuable since they do not rely on specific tools such as temporal parsers which depend themselves on the availability of large-scale annotated data.

9.2 New research directions

We have pointed at the preferences in usage of pronouns and nominal reference in the target language. Currently, most machine translation systems are evaluated on a single reference during training, and are thus optimized towards a single translation possibility. This process may actually be sub-optimal. We have shown corpus statistics that prove that pronouns have several types of translations and are not always translated by pronouns of the same type. Our figures hold according to a single reference. However, they can only be taken as a hint on the real distribution of the preferred usage of pronouns in a language, French in our case. The difference between a grammatical translation and a preferred translation should be further investigated.

In this sense, we think that new resources in the form of multiple reference translations could be created in order to estimate the distribution of the preferred usage of pronouns. As an alternative, we think that the fill-the-gap task introduced by Hardmeier (2014) and used at both shared tasks on cross-lingual pronoun prediction, could be re-used as a ranking task presented to multiple judges. This would produce multiple annotations of a same text and would provide with the means to compute a distribution of (pro)nominal

reference in use.

Concerning our chosen methods to tackle pronoun translation, the value of rule-based systems is difficult to see when they are systematically outperformed by statistical systems. Our experiments were no exception. For our own particular experiments, we saw a problem of under-generation of some of the possible pronoun translations for *it* and *they* but very good pronoun-antecedent matching in the cases covered by the rules. In other words, the system has high precision, but very bad recall. We also concluded from our experiments that contextual features were important predictors for all the classes, but in particular for the classes with high frequency, making a language model good at recall measures. As a matter of fact, a strong language model was the best ranking system in the 2015 shared task on pronoun prediction. Given these two facts, we think that the Its-2 system could be combined with a language model which would compensate specifically for the cases which are not generated. This could be a potential solution for other systems with a similar architecture.

Some latent features have been argued to play a crucial role in pronoun resolution and they could potentially be integrated in the cross-lingual pronoun prediction task. Holler and Irmen (2007), for instance, suggest the information whether the potential antecedent is animate or not, and whether it functions as a topic or not. They also mention the information-structural status of a potential antecedent in terms of providing new or familiar information. When looking at data and thinking about the pronoun resolution process introspectively, these features seem sensible, but they remain difficult to implement. In this sense, we think that better formalization of the existing theoretical linguistic knowledge could be advantageous.

The self-training experiments for the *it* disambiguation task demonstrated the benefit of combining the gold-standard data with noisy data labeled automatically, the silver-standard data. Since the two models have different strengths, in future work we plan to enrich the training data with re-training instances from the silver data where the two systems agree, in order to reduce the amount of noise, following the example of Jiang, Carenini, and R. Ng (2016). Ultimately, we aim at integrating the *it-function* prediction system within a full machine translation pipeline, and into a coreference resolution system.

Furthermore, we think it is worth exploring the anaphoric interpretation of verb tenses more. We have not been able to investigate the effect in machine translation output of

including intersentential verbal relations, but we think this could be even more beneficial than remaining at the sentence level. If, like pronouns, verb tenses are part of a bigger inter-sentential coreference chain, it is reasonable to assume that a consistent translation of the chain would produce better results than the translation of each verb in isolation. The automatically annotated corpus we have provided could be a starting point for a corpus study in this respect.

Last, the field of machine translation is currently at a turning point, with neural models prevailing over phrase-based statistical models. While the possibilities are vast, so is the unknown concerning the two issues treated in this thesis. It is still to see if these two topics are a problem at all for neural models, in particular since these models have access to larger contexts than traditional phrase-based systems, and, if they are, what concrete possibilities these models offer to cope with them. Our self-training experiments with the recurrent neural network classifier and the maximum entropy classifier have given us some clues in this sense. Comparing these two systems, representative of the two different approaches, we saw that instead of one being better than the other, they have complementary strengths.

In this chapter we have concluded our work and presented a review of what we have accomplished in this thesis. We have discussed our specific contributions and findings and have discussed the points which have not been as conclusive as we would have wanted. Throughout all the experiments presented in this thesis, we have aimed at enhancing existing machine translation architectures with linguistic knowledge in ways useful to create more fluent machine translation output. We have looked to understand the translation process better, drawing attention to the discussion of two specific aspects for which there is still room for improvement.

Bibliography

Aarts, Bas (2011). *Oxford Modern English Grammar*. Oxford: Oxford University Press.

Ahern, Aoife (2005). "Mood choice and sentence interpretation in Spanish". In: *Crosslinguistic Views on Tense, Aspect and Modality*. Ed. by Bart Hollebrandse, Angeliek van Hout, and Co Vet. Vol. 13. Amsterdam: Editions Rodipi B.V., pp. 201–214.

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.

— (2005). "Troubles on the Right Frontier". In: *Proceedings of Constraints in Discourse 2005*. Amsterdam Philadelphia: John Benjamins Publishing Company, pp. 29–52.

Asher, Nicholas, Pascal Denis, Jonas Kuhn, Erik Larson, Eric McCready, Alexis Palmer, Brian Reese, and Linton Wang (2004). "Extracting and Using Discourse Structure to Resolve Anaphoric Dependencies: Combining Logico-Semantic and Statistical Approaches". In: *Proceedings of the 11th Conference Traitement Automatique du Langage Naturel, Workshop SDRT*. TALN 2004. Fès.

Asher, Nicholas and Alex Lascarides (1995). "Lexical Disambiguation in a Discourse Context". In: *Journal of Semantics* 12.1, pp. 69–108.

Bar-Hillel, Yoshua (1960). "The Present Status of Automatic Translation of Languages". In: *Advances in Computers* 1, pp. 91–163.

Bawden, Rachel (2016). "Cross-lingual Pronoun Prediction with Linguistically Informed Features". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 564–570.

Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra (1996). "A Maximum Entropy Approach to Natural Language Processing". In: *Computational Linguistics* 22.1, pp. 39–71.

Bergsma, Shane and David Yarowsky (2011). "NADA: A Robust System for Non - referential Pronoun Detection". In: *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*. Ed. by Iris Hen-

drickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov. Lecture Notes in Artificial Intelligence. Faro, Portugal: Springer, pp. 12–23.

Bhat, Darbhe Narayana Shankara (2004). *Pronouns*. Oxford: Oxford University Press.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič (2013). "Joint morphological and syntactic analysis for richly inflected languages". In: *Transactions of the Association for Computational Linguistics* 1, pp. 415–428.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (2013). "Findings of the 2013 Workshop on Statistical Machine Translation". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. WMT13. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1–44.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna (2014). "Findings of the 2014 Workshop on Statistical Machine Translation". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. WMT14. Baltimore, Maryland: Association for Computational Linguistics, pp. 12–58.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi (2015). "Findings of the 2015 Workshop on Statistical Machine Translation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. WMT15. Lisbon, Portugal: Association for Computational Linguistics, pp. 1–46.

Brennan, Susan E., Marilyn W. Freidman, and Carl J. Pollard (1987). "A Centering Approach to Pronouns". In: *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*. ACL 1987. Stanford, California: Association for Computational Linguistics, pp. 155–162.

Bresnan, Joan (2001). *Lexical Functional Grammar*. Oxford: Blackwell Publishers.

Büring, Daniel (2005). *Binding Theory*. New York: Cambridge University Press.

Callin, Jimmy, Christian Hardmeier, and Jörg Tiedemann (2015). "Part-of-Speech Driven Cross-Lingual Pronoun Prediction with Feed-Forward Neural Networks". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 59–64.

Cettolo, Mauro, Christian Girardi, and Marcello Federico (2012). "WIT^3: Web Inventory of Transcribed and Translated Talks". In: *Proceedings of the 16th Conference of the European Association for Machine Translation*. EAMT 2012. Trento, Italy, pp. 261–268.

Chen, Stanley F. and Joshua Goodman (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Tech. rep. Cambridge, MA: Computer Science Group, Harvard University.

Chollet, François (2015). *Keras*. `https://github.com/fchollet/keras`.

Chomsky, Noam (1972). *Syntactic Structures*. Mouton.

— (1980). "On Binding". In: *Linguistic Inquiry* 11.1, pp. 1–46.

— (1981). *Lectures on Government and Binding: The Pisa Lectures*. Mouton de Gruyter.

— (1995). *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.

Chrupała, Grzegorz, Georgiana Dinu, and Josef van Genabith (2008). "Learning Morphology with Morfette". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. LREC 2008. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 2362–2367.

Clark, Herbert H. (1975). "Bridging". In: *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*. TINLAP 1975. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 169–174.

Clements, Joseph Clancy (2008). "El español a través de la lingüística". In: ed. by Jennifer D. Ewald and Anne Edstrom. Somerville, MA: Cascadilla Press. Chap. Me dicen que suena raro cuando digo yo en todo momento: ¿por qué no es necesario usar el pronombre?, pp. 83–94.

Comrie, Bernard (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Culicover, Peter and Ray Jackendoff (2005). *Simpler Syntax*. New York: Oxford University Press.

Dabre, Raj, Yevgeniy Puzikov, Fabien Cromieres, and Sadao Kurohashi (2016). "The Kyoto University Cross-Lingual Pronoun Translation System". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 571–575.

Dahl, Östen and Viveka Velupillai (2013). "Tense and Aspect". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath.

Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: http://wals.info/chapter/s7.

De Beaugrande, Robert and Wolfgang Dressler (1981). *Introduction to Text Linguistics*. Essex: Longman Linguistics Library.

Declerck, Renaat (2007). "Distinguishing between the aspectual categories '(a) telic', '(im) perfective', and '(non) bounded'". In: *Kansas Working Papers in Linguistics* 29, pp. 48–64.

Denkowski, Michael and Alon Lavie (2011). "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*. WMT11. Edinburgh, Scotland: Association for Computational Linguistics, pp. 85–91.

Depraetere, Ilse (1995). "On the necessity of distinguishing between (un)boundedness and (a)telicity". In: *Linguistics and Philosophy* 18.1, pp. 1–19.

Dipper, Stefanie, Melanie Seiss, and Heike Zinsmeister (2012). "The Use of Parallel and Comparable Data for Analysis of Abstract Anaphora in German and English". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation*. LREC 2012. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 138–145.

Dipper, Stefanie and Heike Zinsmeister (2010). "Towards a Standard for Annotating Abstract Anaphora". In: *Proceedings of the LREC Workshop on Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments*. LREC10-W4. Valleta, Malta: European Language Resources Association (ELRA), pp. 54–59.

Eijck, Jan van and Hans Kamp (1997). "Representing Discourse in Context". In: *Handbook of logic and language*. Ed. by Johan van Benthem and Alice ter Meulen. Amsterdam: Elsevier Science B.V., pp. 179–237.

Eisele, Andreas and Yu Chen (2010). "MultiUN: A Multilingual Corpus from United Nation Documents". In: *Proceedings of the 7th International Conference on Language Resources and Evaluation*. LREC 2010. Valletta, Malta: European Language Resources Association (ELRA), pp. 2868–2872.

Engelberg, Stefan (1999). ""Punctuality" and Verb Semantics". In: *University of Pennsylvania Working Papers in Linguistics* 6 (1), pp. 1–16.

Evans, Richard (2001). "Applying Machine Learning Toward an Automatic Classification of IT". In: *Literary and Linguistic Computing* 16.1, pp. 45–57.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin (2008). "LIBLINEAR: A Library for Large Linear Classification". In: *Journal of Machine Learning Research* 9, pp. 1871–1874.

Ferrández, Antonio and Jesús Peral (2000). "A Computational Approach to Zero Pronouns in Spanish". In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. ACL 2000. Hong Kong, pp. 166–172.

Friedrich, Annemarie and Alexis Palmer (2014). "Automatic Prediction of Aspectual Class of Verbs in Context". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2014. Baltimore, Maryland: Association for Computational Linguistics, pp. 517–523.

Garnham, Alan (2001). *Mental Models and the Interpretation of Anaphora*. Sussex: Psychology Press.

Ghorbel, Hatem, Afzal Ballim, and Giovanni Coray (2001). "ROSETTA: Rhetorical and semantic environment for text alignment". In: *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK, pp. 224–233.

Gojun, Anita and Alexander Fraser (2012). "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 2012. Avignon, France, pp. 726–735.

Gong, Zhengxian, Min Zhang, Chewlim Tan, and Guodong Zhou (2012a). "Classifier-based Tense Models for SMT". In: *Proceedings of the 25th International Conference on Computational Linguistics*. COLING 2012. Mumbai, India: The COLING 2012 Organizing Committee, pp. 411–420.

— (2012b). "N-gram-based tense models for statistical machine translation". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012. Jeju Island, Korea: Association for Computational Linguistics, pp. 276–285.

Grishman, Ralph and Beth Sundheim (1995). "Design of the MUC-6 Evaluation". In: *Sixth Message Understanding Conference*. MUC-6. Columbia: Association for Computational Linguistics, pp. 1–11.

Grisot, Cristina (2015). "Temporal reference: empirical and theoretical perspectives. Converging evidence from English and Romance". PhD thesis. Geneva, Switzerland: Université de Genève.

Grisot, Cristina and Bruno Cartoni (2012). "Une description bilingue des temps verbaux: étude contrastive en corpus". In: *Nouveaux cahiers de linguistique française* 30, pp. 101–117.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (1986). *Towards a Computational Theory of Discourse Interpretation*. Preliminary Draft - Never published.

— (1995). "Centering: A Framework for Modelling the Local Coherence of Discourse". In: *Computational Linguistics* 2.21, pp. 203–225.

Guillou, Liane (2012). "Improving Pronoun Translation for Statistical Machine Translation". In: *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 1–10.

— (2015). "Automatic Post-Editing for the DiscoMT Pronoun Translation Task". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 65–71.

— (2016). "Incorporating Pronoun Function into Statistical Machine Translation". PhD thesis. Scotland, UK: University of Edinburgh.

Guillou, Liane, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis (2016). "Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 525–542.

Guillou, Liane, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber (2014). "ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3191–3198.

Haegeman, Liliane (1994). *Introduction to Government and Binding Theory*. Oxford: Blackwell Publishers.

Halliday, Makh R. and Ruqaiya Hasan (1976). *Cohesion in English*. New York: Longman Inc.

Harabagiu, Sanda M. and Steven J. Maiorano (1999). "Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence". In: *Proceedings of*

The Relation of Discourse/Dialogue Structure and Reference Workshop. Maryland, USA: Association for Computational Linguistics, pp. 29–38.

Hardmeier, Christian (2014). "Discourse in Statistical Machine Translation". PhD thesis. Uppsala, Sweden: Department of Linguistics and Philology, Uppsala University.

— (2015). "A Document-Level SMT System with Integrated Pronoun Prediction". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 72–77.

— (2016). "Pronoun Prediction with Latent Anaphora Resolution". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 576–580.

Hardmeier, Christian and Marcello Federico (2010). "Modelling Pronominal Anaphora in Statistical Machine Translation". In: *Proceedings of the 7th International Workshop on Spoken Language Translation*. IWSLT 2010. Paris, France, pp. 283–289.

Hardmeier, Christian, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo (2015). "Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal, pp. 1–16.

Hardmeier, Christian, Joakim Nivre, and Jörg Tiedemann (2012). "Document-Wide Decoding for Phrase-Based Statistical Machine Translation". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012. Jeju Island, Korea: Association for Computational Linguistics, pp. 1179–1190.

Hardmeier, Christian, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely (2016). *DiscoMT 2015 Shared Task on Pronoun Translation*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

Hardmeier, Christian, Jörg Tiedemann, and Joakim Nivre (2013). "Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. Seattle, Washington: Association for Computational Linguistics, pp. 380–391.

Heafield, Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. WMT11. Edinburgh, UK: Association for Computational Linguistics, pp. 187–197.

Henderson, James, Paola Merlo, Gabriele Musillo, and Ivan Titov (2008). "A Latent Variable Model of Synchronous Parsing for Syntactic and Semantic Dependencies". In: *Proceedings of the 12th Conference on Computational Natural Language Learning*. CONLL 2008. Manchester, UK, pp. 178–182.

Herrmann, Teresa, Jan Niehues, and Alex Waibel (2015). "Source Discriminative Word Lexicon for Translation Disambiguation". In: *Proceedings Proceedings of the International Workshop on Spoken Language Translation*. IWSLT 2015. Da Nang, Vietnam, pp. 135–142.

Hobbs, Jerry (1978). "Resolving Pronoun References". In: *Lingua* 1.44, pp. 311–338.

Holler, Anke and Lisa Irmen (2007). "Empirically Assessing Effects of the Right Frontier Constraint". In: *DAARC*. Ed. by António Horta Branco. Lecture Notes in Computer Science. Springer, pp. 15–27.

Hutchins, John (2010). "Machine translation: a concise history". In: *Journal of Translation Studies. Special issue: The teaching of computer aided translation* 13.1–2, pp. 29–70.

Jiang, Kailang, Giuseppe Carenini, and Raymond Ng (2016). "Training Data Enrichment for Infrequent Discourse Relations". In: *Proceedings the 26th International Conference on Computational Linguistics: Technical Papers*. COLING 2016. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 2603–2614.

Joty, Shafiq, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov (2014). "DiscoTK: Using Discourse Structure for Machine Translation Evaluation". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. WMT14. Baltimore, Maryland: Association for Computational Linguistics, pp. 402–408.

Kamp, Hans and Uwe Reyle (1993). *From Discourse to Logic. Introduction to Modeltheorectic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Kehler, Andrew, Douglas Appelt, Lara Taylor, and Aleksandr Simma (2004). "The (Non) Utility of Predicate-Argument Frequencies for Pronoun Interpretation". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. HLT-NAACL 2004. Boston: Association for Computational Linguistics, pp. 289–296.

Kennedy, Christopher and Banimir Boguraev (1996). "Anaphora for everyone: Pronominal anaphora resolution without a parser". In: *In Proceedings of 16th International Conference on Computational Linguistics*. COLING 1996. Copenhagen: John Wiley and Sons, Ltd.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer (2008). "A large scale classification of English Verbs". In: *Language Resources and Evaluation* 42.1, pp. 21–40.

Klappholtz, David and Abe Lockman (1975). "Contextual Reference Resolution". In: *Proceedings of the 13th Annual Meeting of the Association for Computational Linguistics*. ACL 1975. Minnesota: Association for Computational Linguistics, pp. 4–25.

Klenner, Manfred, Angela Fahrni, and Rico Sennrich (2010). "Real Anaphora Resolution is Hard. The case of German". In: *Text, Speech and Dialogue: Proceedings of the 13th International Conference TSD*. Vol. 6231. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, pp. 109–116.

Klenner, Manfred, Don Tuggener, Angela Fahrni, and Rico Sennrich (2010). "Anaphora Resolution with Real Processing". In: *Lecture Notes in Computer Science*. Lecture Notes in Computer Science 6233, pp. 215–225.

Koehn, Philipp (2004). "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004*. EMNLP 2004. Barcelona, Spain, pp. 388–395.

— (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of the 10th Machine Translation Summit*. MT Summit X. Phuket, Thailand, pp. 79–86.

— (2010). *Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp and Hieu Hoang (2007). "Factored Translation Models". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2007. Prague, Czech Republic, pp. 868–876.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoli, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings*

of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. ACL 2007. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

Kolhatkar, Varada (2015). "Resolving Shell Nouns". PhD thesis. Toronto, Canada: University of Toronto.

Kučová, Lucie and Eva Hajičová (2005). "Coreferential Relations in the Prague Dependency Treebank". In: *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution 2004*. DAARC. San Miguel, Azores, pp. 97–102.

Lappin, Shalom and Herbert J. Leass (1994). "An Algorithm for Pronominal Anaphora Resolution". In: *Computational Linguistics* 20.4, pp. 535–561.

Lascarides, Alex and Nicholas Asher (1993). "Temporal interpretation, discourse relations and commonsense entailment". In: *Linguistics and Philosophy* 16.5, pp. 37–493.

— (2007). "Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure". In: *Computing Meaning: Volume 3*. Kluwer Academic Publishers, pp. 87–124.

Lascarides, Alex, Nicholas Asher, and Jon Oberlander (1992). "Inferring Discourse Relations in Context". In: *Proceedings of the 30th Annual Meeting of the Association of Computational Linguistics*. ACL 1992. Newark, Delaware, pp. 1–8.

Laurent, Dominique (2001). *De la résolution des anaphores*. Tech. rep. Synapse Développement.

Le Nagard, Ronan and Philipp Koehn (2010). "Aiding Pronoun Translation with Co-Reference Resolution". In: *Proceedings of the Joint 5th Workshop on Statistical Machine Translation*. WMT10. Uppsala, Sweden: Association for Computational Linguistics, pp. 258–267.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2011). "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. CONLL 2011. Portland, Oregon: Association for Computational Linguistics, pp. 28–34.

Lee, Timothy, Alex Lutz, and Jinho D. Choi (2016). "QA-It: Classifying Non-Referential It for Question Answer Pairs". In: *Proceedings of the ACL 2016 Student Research*

Workshop. Berlin, Germany: Association for Computational Linguistics, pp. 132–137.

Loáiciga, Sharid (2013). "Résolution d'anaphores et traitement des pronoms en traduction automatique à base de règles". In: *Proceedings of the 20th Conference Traitement Automatique du Langage Naturel*. TALN 2013. Les Sables-d'Olonne, pp. 683–690.

— (2015). "Predicting Pronoun Translation Using Syntactic, Morphological and Contextual Features from Parallel Data". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 78–85.

Loáiciga, Sharid and Cristina Grisot (2016). "Predicting and Using a Pragmatic Component of Lexical Aspect of Simple Past Verbal Tenses for Improving English-to-French Machine Translation". In: *Linguistic Issues in Language Technology* 13.3, pp. 1–36.

Loáiciga, Sharid, Liane Guillou, and Christian Hardmeier (SUBMITTED). "What is it?: Disambiguating the different readings of the pronoun 'it'". In: SUBMITTED. SUBMITTED. submitted.

— (2016). "It-disambiguation and source-aware language models for cross-lingual pronoun prediction". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 581–588.

Loáiciga, Sharid, Thomas Meyer, and Andrei Popescu-Belis (2014). "English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. LREC 2014. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 674–681.

Loáiciga, Sharid and Éric Wehrli (2015). "Rule-Based Pronominal Anaphora Treatment for Machine Translation". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 86–93.

Lopatková, Markéta, Martin Plátek, and Petr Sgall (2008). "Functional Generative Description, Restarting Automata and Analysis by Reduction". In: *Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5.*

Ed. by Franc Marušič and Rok Žaucer. Vol. 19. Frankfurt am Main: Peter Lang GmbH, pp. 173–190.

Luong, Ngoc Quang, Lesly Miculicich Werlen, and Andrei Popescu-Belis (2015). "Pronoun Translation and Prediction with or without Coreference Links". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 94–100.

Luong, Ngoc Quang and Andrei Popescu-Belis (2016a). "A contextual language model to improve machine translation of pronouns by re-ranking translation hypothesis". In: *Baltic Journal of Modern Computing - EAMT 2016* 2, pp. 292–304.

— (2016b). "Pronoun Language Model and Grammatical Heuristics for Aiding Pronoun Prediction". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 589–595.

Luotolahti, Juhani, Jenna Kanerva, and Filip Ginter (2016). "Cross-Lingual Pronoun Prediction with Deep Recurrent Neural Networks". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 596–601.

Mann, William C. and Sandra A. Thompson (1987). "Rhetorical Structure Theory: A Framework for the Analysis of Texts". In: *Papers in Pragmatics* 1, pp. 79–105.

— (1988). "Rhetorical Structure Theory: Towards a functional theory of text organization". In: *Text* 8.3, pp. 243–281.

Manning, Christopher and Dan Klein (2003). *MaxEnt Models, Conditional Estimation, and Optimization without Magic*. Tutorial at HLT-NAACL and 41st ACL conferences. URL: `https://people.eecs.berkeley.edu/~klein/papers/maxent-tutorial-slides.pdf`.

Marcu, Daniel, Lynn Carlson, and Maki Watanabe (2000). "The Automatic Translation of Discourse Structures". In: *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. NAACL 2000. Seattle, Washington: Association for Computational Linguistics, pp. 9–17.

Martschat, Sebastian and Michael Strube (2015). "Latent Structures for Coreference Resolution". In: *Transactions of the Association for Computational Linguistics* 3, pp. 405–418.

Mauser, Arne, Saša Hasan, and Hermann Ney (2009). "Extending Statistical Machine Translation with Discriminative and Trigger-based Lexicon Models". In: *Proceed-*

ings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. EMNLP 2009. Singapore: Association for Computational Linguistics, pp. 210–218.

Meyer, Thomas (2014). "Discourse-level Features for Statistical Machine Translation". PhD thesis. Lausanne, Switzerland: École Polytechnique Fédérale de Lausanne.

Meyer, Thomas, Cristina Grisot, and Andrei Popescu-Belis (2013). "Detecting Narrativity to Improve English to French Translation of Simple Past Verbs". In: *Proceedings of the First DiscoMT Workshop at the 51th Annual Meeting of the Association for Computational Linguistics*. DiscoMT 2013. Sofia, Bulgaria, pp. 33–42.

Meyer, Thomas and Andrei Popescu-Belis (2012). "Using Sense-labeled Discourse Connectives for Statistical Machine Translation". In: *Proceedings of the Workshop on Hybrid Approaches to Machine Translation at EACL 2012*. HyTra. Avignon, France, pp. 129–138.

Meyer, Thomas, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo (2012). "Machine Translation of Labeled Discourse Connectives". In: *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*. AMTA 2012.

Meyer, Thomas, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni (2011). "Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation". In: *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL 2011. Portland, Oregon: Association for Computational Linguistics, pp. 194–203.

Mitkov, Ruslan (2002). *Anaphora Resolution*. Harlow: Pearson Education Limited.

— (2003). In: *The Oxford Handbook of Computational Linguistics*. Ed. by Ruslan Mitkov. Oxford: Oxford University Press. Chap. Anaphora Resolution, pp. 266–285.

— (2010). "Discourse Processing". In: *Computational Linguistics and Natural Language Handbook*. Ed. by Chris Fox Alexander Clark and Shalom Lappin. West Sussex: Wiley-Blackwell Publishers, pp. 599–629.

Mitkov, Ruslan and Catalina Barbu (2002). "Using Bilingual Corpora to Improve Pronoun Resolution". In: *Languages in Contrast* 4.2, pp. 201–211.

Mitkov, Ruslan, Richard Evans, and Constantin Orăsan (2002). "A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method". In: *Pro-*

ceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics. CICLing 2000. Mexico City, pp. 168–186.

Moens, Marc and Mark Steedman (1988). "Temporal Ontology and Temporal Reference". In: *Computational Linguistics* 14.2, pp. 15–28.

Moeschler, Jacques and Anne Reboul (1994). *Dictionnaire encyclopédique de pragmatique*. Paris: Éditions du Seuil.

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme (2012). "Annotated Gigaword". In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. AKBC-WEKEX. Montreal, Canada: Association for Computational Linguistics, pp. 95–100.

Navarretta, Costanza (2004). "Resolving individual and abstract anaphora in texts and dialogues". In: *Proceedings of the 20th International Conference on Computational Linguistics*. COLING 2004. Geneva, Switzerland: Association for Computational Linguistics, pp. 233–239.

Nedoluzhko, Anna, Jiří Mírovský, and Michal Novák (2013). "A Coreferentially Annotated Corpus and Anaphora Resolution for Czech". In: *Computational Linguistics and Intellectual Technologies*. Moskva, Russia: ABBYY, pp. 467–475.

Neeleman, Ad and Kriszta Szendői (2005). "Pro Drop and Pronouns". In: *Proceedings of the 24th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 299–307.

— (2007). "Radical Pro Drop and the Morphology of Pronouns". In: *Linguistic Inquiry* 38.4, pp. 671–714.

Ng, Vincent (2010). "Supervised Noun Phrase Coreference Research: The First Fifteen Years". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL 2010. Uppsala, Sweden, pp. 1396–1411.

Novák, Michal (2011). "Utilization of Anaphora in Machine Translation". In: *Proceedings of the 20th Annual Conference of Doctoral Students–Contributed Papers: Part I*. WDS11. Prague: Matfyzpress, pp. 155–160.

— (2016). "Pronoun Prediction with Linguistic Features and Example Weighing". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 602–608.

Novák, Michal, Anna Nedoluzhko, and Zdeněk Žabokrtský (2013). "Translation of "It" in a Deep Syntax Framework". In: *Proceedings of the Workshop on Discourse in*

Machine Translation. DiscoMT 2015. Sofia, Bulgaria: Association for Computational Linguistics, pp. 51–59.

Och, Franz Josef (2003). "Minimum Error Rate Training in Statistical Machine Translation". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL 2003. Sapporo, Japan: Association for Computational Linguistics, pp. 160–167.

Och, Franz Josef and Hermann Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". In: *Computational Linguistics* 29, pp. 19–51.

Olsen, Mari, David Traum, Carol Van Ess-Dykema, and Amy Weinberg (2001). *Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System*. Tech. rep. LAMP-TR-070,CS-TR-4248,UMIACS-TR-2001-33. University of Maryland, College Park.

Palomar, Manuel, Lidia Moreno, Jesús Peral, Rafael Muñoz, Antonio Ferrández, Patricio Martínez Barco, and Maximiliano Saiz Noeda (2001). "An algorithm for anaphora resolution in Spanish texts". In: *Computational Linguistics* 27 (4), pp. 545–567.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL 2002. Philadelphia: Association for Computational Linguistics, pp. 311–318.

Partee, Barbara Hall (1973). "Some Structural Analogies between Tenses and Pronouns in English". In: *The Journal of Philosophy* 70.18, pp. 601–609.

— (1984). "Nominal and Temporal Anaphora". In: *Linguistics and Philosophy* 7.3, pp. 243–286.

Pham, Ngoc-Quan and Lonneke van der Plas (2015). "Predicting Pronouns across Languages with Continuous Word Spaces". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 101–107.

Pineda, Luis and Ivan Meza (2006). "The Spanish Pronominal Clitic System". In: *Procesamiento del lenguaje natural* 34, pp. 67–103.

Poesio, Massimo, Simone Paolo Ponzetto, and Yannick Versley (2010). "Computational Models of Anaphora Resolution: A Survey". URL: `http://cswww.essex.ac.uk/poesio/papers.html`.

Poesio, Massimo and Renata Vieira (1998). "A corpus Based Investigation of Definite Description Use". In: *Computational Linguistics* 24.2, pp. 183–216.

Polanyi, Livia (1988). "A formal model of the structure of discourse". In: *Journal of Pragmatics* 12.5, pp. 601–638.

Popescu-Belis, Andrei, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey (2012). "Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation*. LREC 2012. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2716–2720.

Princeton University (2010). *WordNet*. URL: `http://wordnet.princeton.edu`.

Qiu, Long, Min-Yen Kan, and Tat-Seng Chua (2004). "A Public Reference Implementation of the RAP Anaphora Resolution Algorithm". In: *Proceedings of Language Resources and Evaluation Conference*. LREC 2004. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 291–294.

Ramm, Anita, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser (2017). "Annotating tense, mood and voice for English, French and German". In: *Proceedings of 55th annual meeting of the Association for Computational Linguistics*. ACL17. Software Demonstration. Vancouver, Canada: Association for Computational Linguistics.

Recasens, Marta and M. Antònia Martí (2010). "AnCora-Co: Coreferentially annotated corpora for Spanish and Catalan". In: *Language Resources and Evaluation* 44.4, pp. 315–345.

Reichenbach, Hans (1947). *Elements of symbolic logic*. New York: Mcmillan.

Reinhart, Tanya (1983). *Anaphora Resolution and Semantic Interpretation*. London: Croom Helm.

Rello, Luz and Iustina Ilisei (2009). "A Rule-Based Approach to the Identification of Spanish Zero Pronouns". In: *Proceedings of the Student Research Workshop in the Conference on Recent Advances in Natural Language Processing*. RANLP-09. Borovets, Bulgaria, pp. 60–65.

Rello, Luz, Pablo Suárez, and Ruslan Mitkov (2010). "A machine learning method for identifying impersonal constructions and zero pronouns in Spanish". In: *Procesamiento del Lenguaje Natural* 45, pp. 281–285.

Russo, Lorenza, Yves Scherrer, Jean-Philippe Goldman, Sharid Loáiciga, Luka Nerima, and Éric Wehrli (2011). "Étude interlangues de la distribution et des ambiguïtés syntaxique des pronoms". In: *Proceedings of the 18th Conference Traitement Automatique du Langage Naturel*. TALN 2011. Montpellier.

Sagot, Benoît (2010). "The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. LREC 2010. Valletta, Malta: European Language Resources Association (ELRA), pp. 2744–2751.

Scherrer, Yves, Lorenza Russo, Jean-Philippe Goldman, Sharid Loáiciga, Luka Nerima, and Éric Wehrli (2011). "La traduction automatique des pronoms: Problèmes et perspectives". In: *Proceedings of the 18th Conference Traitement Automatique du Langage Naturel*. TALN 2011. Montpellier.

Schmid, Helmut (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Scrouton, Gordon (2011). *What is Cohesion & Coherence?* URL: `http://gordonscruton.blogspot.se/2011/08/what-is-cohesion-coherence-cambridge.html`.

Silva, Lucia (2010). "Fine-Tuning in Brazilian Portuguese-English Statistical Transfer Machine Translation: Verbal Tenses". In: *Proceedings of the NAACL-HLT 2010 Student Research Workshop*. Los Angeles, CA: Association for Computational Linguistics, pp. 58–63.

Somers, Harold (2003). In: *The Oxford Handbook of Computational Linguistics*. Ed. by Ruslan Mitkov. Oxford: Oxford University Press. Chap. Machine Translation: Latest Developments, pp. 512–528.

Song, Xingyi, Trevor Cohn, and Lucia Specia (2013). "BLEU deconstructed: Designing a Better MT Evaluation Metric". In: *International Journal of Computational Linguistics and Applications* 4.2. Ed. by Alexander Gelbukh, pp. 29–44.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". In: *Computational Linguistics* 27.4, pp. 521–544.

Sparck Jones, Karen (1994). "Natural Language Processing: A Historical Review". In: *Current Issues in Computational Linguistics: In Honour of Don Walker*. Ed. by Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer. Vol. 9. Linguistica Computazionale. Amsterdam: Springer, pp. 3–16.

Stede, Manfred (2012). *Disourse Processing*. Toronto: Morgan and Claypool Publishers.

Stolcke, Andreas, Jing Zheng, Wen Wang, and Victor Abrash (2011). "SRILM at Sixteen: Update and Outlook". In: *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. Waikoloa, Hawaii.

Stoyanov, Veselin, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom (2009). "Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. ACL-IJCNLP 2009. Suntec, Singapore: Association for Computational Linguistics, pp. 656–664.

— (2010). "Coreference Resolution with Reconcile". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL 2010. Uppsala, Sweden: Association for Computational Linguistics, pp. 156–161.

Strube, Michael (2007). "Corpus-based and Machine Learning Approaches to Coreference Resolution". In: *Anaphors in Text. Cognitive, Formal and Applied Approaches to Anaphoric Reference*. Ed. by Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees. Amsterdam Philadelphia: John Benjamins Publishing Company, pp. 207–222.

Stymne, Sara (2016). "Feature Exploration for Cross-Lingual Pronoun Prediction". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 609–615.

Tiedemann, Jörg (2015). "Baseline Models for Pronoun Prediction and Pronoun-Aware Translation". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 108–114.

— (2016). "A Linear Baseline Classifier for Cross-Lingual Pronoun Prediction". In: *Proceedings of the First Conference on Machine Translation*. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 616–619.

Vendler, Zeno (1957). "Verbs and Times". In: *The Philosophical Review* 66.2, pp. 143–160.

Versley, Yannick, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti (2008). "BART: A Modular Toolkit for Coreference Resolution". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*. HLT-Demonstrations 2008. Columbus, Ohio: Association for Computational Linguistics, pp. 9–12.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney (2006). "Error Analysis of Statistical Machine Translation Output". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. LREC 2006. Genoa, Italy: European Language Resources Association (ELRA), pp. 697–702.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda (2005). *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium. LDC2006T06.

Webber, Bonnie (1979). *A Formal Approach to Discourse Anaphora*. New York: Garland Publishing Inc.

— (1990). *Structure and Ostension in the Interpretation of Discourse Deixis*. Tech. rep. MS-CIS-90-58. University of Pennsylvania.

— (2014). *Discourse and SMT*. Presentation at the Ninth MT Marathon, Trento. URL: `http://www.statmt.org/mtm14/uploads/Main/DiscourseSMT_MTM2014.pdf`.

Webber, Bonnie, Markus Egg, and Valia Kordoni (2011). "Discourse Structure and Language Technology". In: *Natural Language Engineering* 18.4, pp. 437–490.

Webber, Bonnie and Aravind K. Joshi (2012). "Discourse Structure and Computation: Past, Present and Future". In: *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. ACL 2012. Jeju Island: Association for Computational Linguistics, pp. 42–54.

Webber, Bonnie, Matthew Stone, Aravind K. Joshi, and Alistair Knott (2003). "Anaphora and Discourse Structure". In: *Computational Linguistics* 29.4, pp. 545–587.

Wehrli, Éric (2007). "Fips, a "Deep" Linguistic Multilingual Parser". In: *Proceedings of the Workshop on Deep Linguistic Processing*. Prague, Czech Republic: Association for Computational Linguistics, pp. 120–127.

Wehrli, Éric and Luka Nerima (2009). "L'analyseur syntaxique Fips". In: *Proceedings of the ATALA Workshop at the 11th Conference on Parsing Technologies*. IWPT 2009. Paris, France, pp. 1–8.

Wehrli, Éric, Luka Nerima, and Yves Scherrer (2009). "Deep Linguistic Multilingual Translation and Bilingual Dictionaries". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. WMT09. Athens, Greece, pp. 90–94.

Weiner, Jochen Stefan (2014). "Pronominal Anaphora in Machine Translation". Master of Science. Karlsruhe Institute of Technology.

Wetzel, Dominikus (2016). "Cross-lingual Pronoun Prediction for English, French and German with Maximum Entropy Classification". In: *Proceedings of the First Con-*

ference on Machine Translation. WMT16. Berlin, Germany: Association for Computational Linguistics, pp. 620–626.

Wetzel, Dominikus, Adam Lopez, and Bonnie Webber (2015). "A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. DiscoMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 115–121.

Ye, Yang, Victoria Li Fossum, and Steven Abney (2006). "Latent Features in Automatic Tense Translation between Chinese and English". In: *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia: Association for Computational Linguistics, pp. 48–55.

Ye, Yang, Karl-Michael Schneider, and Steven Abney (2007). "Aspect Marker Generation for English-to-Chinese Machine Translation". In: *Proceedings of the Eleventh Machine Translation Summit*. MT SUMMIT XI. Copenhagen, Danmark, pp. 521–527.

Yllescas, Juan Carlos Tordera (2012). "Propuesta de traducción sintáctico-semántica: el tratamiento anáforico a través de la LFG y la SDRT". In: *Procesamiento del Lenguaje Natural* 48, pp. 13–20.

Zhang, Ying, Stephan Vogel, and Alex Waibel (2004). "Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*. LREC 2004. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 2051–2054.

Zufferey, Sandrine and Jacques Moeschler (2012). *Initiation à l'étude du sens: sémantique et pragmatique*. Auxerre: Sciences Humaines Editions.

Appendix A

Appendix

Algorithm 1 Pseudocode for VP identification

```
1: procedure IDENTIFYVPS(sentence)
2:     initialization: VP ← [], VPs ← [], Sentence ← Parsed EN sentence in CoNLL format,
   POStag ← Sentence[2], Head ← Sentence[3], DependencyRelation ← Sentence[4]
3:     for word in sentence do
4:         if POStag in ['MD', 'VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ', 'RP'] then
5:             if VP = [] then
6:                 Append Word to VP
7:             else if Head in VP and DependencyRelation in ['VC', 'PRT'] then
8:                 Append Word to VP
9:             else
10:                Append VP to VPs
11:                VP ← []
12:                Append Word to VP
13:            end if
14:        end if
15:    end for
16:    Return VPs
17: end procedure
```

Algorithm 2 Pseudocode for English tense and voice determination

```
1: procedure LABEL VPS(List of VPs in a sentence)
2:     VPs ← []
3:     POSsequence ← concatenated POS-tags of VP's constituting words
4:     FutureModals ← ['will', 'shall']
5:     ConditionalModals ← ['should', 'could', 'would', 'ought', 'must', 'can', 'may',
   'might']
6:     for each VP in VPs do
7:         if POSsequence in ['VBD VBN', 'VBD RP VBN', 'VBD VB VBN', 'VBD
   VB VBN RP'] and VP[0][1][1] in ['was', 'were'] then
8:             tense ← 'simple_past'; voice ← 'passive'
9:         else if POSsequence in ['VBD VBG VBN', 'VBD VBG RP VBN'] then
10:            tense ← 'past_continuous'; voice ← 'passive'
11:        else if POSsequence in ['VBD VBN VBN', 'VBD VBN VBN RP'] then
12:            tense ← 'past_perfect'; voice ← 'passive'
13:        else if POSsequence in ['VBZ VBN', 'VBZ VBN RP', 'VBP VBN', 'VBP
   VBN RP'] and VP[0][1] in ['is', 'are'] then
14:            tense ← 'present'; voice ← 'passive'
15:        else if POSsequence in ['VBZ VBG VBN', 'VBZ VBG VBN RP', 'VBP
   VBG VBN', 'VBP VBG VBN RP'] then
16:            tense ← 'present_continuous'; voice ← 'passive'
17:        else if POSsequence in ['VBZ VBN VBN', 'VBZ VBN VBN RP', 'VBP
   VBN VBN', 'VBP VBN VBN RP'] then
18:            tense ← 'present_perfect'; voice ← 'passive'
```

```
19:         else if POSsequence in ['MD VB VBN', 'MD VB VBN RP' ] and VP[1][1]
    = 'be' then
20:             if VP[0][1] in FutureModals  then
21:                 tense ← 'future'; voice ← 'passive'
22:             else if VP[0][1] in ConditionalModals then
23:                 tense ← 'conditional'; voice ← 'passive'
24:             end if
25:         else if POSsequence in ['MD VB VBG VBN', 'MD VB VBG VBN RP']
    then
26:             if VP[0][1] in FutureModals  then
27:                 tense ← 'future_continuous'; voice ← 'passive'
28:             else if VP[0][1] in ConditionalModals then
29:                 tense ← 'conditional_continuous'; voice ← 'passive'
30:             end if
31:         else if POSsequence in ['MD VB VBN VBN', 'MD VB VBN VBN RP']
    then
32:             if VP[0][1] in FutureModals then
33:                 tense ← 'future_perfect'; voice ← 'passive'
34:             else if VP[0][1] in ConditionalModals then
35:                 tense ← 'conditional_perfect'; voice ← 'passive'
36:             end if
37:         else if POSsequence in ['VBD', 'VBD RP', 'VBD VB', 'VBD VB RP'] then
38:             tense ← 'simple_past'; voice ← 'active'
39:         else if POSsequence in ['VBD VBG', 'VBD VBG RP'] then
40:             tense ← 'past_continuous'; voice ← 'active'
41:         else if POSsequence in ['VBD VBN', 'VBD VBN RP'] then
42:             tense ← 'past_perfect'; voice ← 'active'
43:         else if POSsequence in ['VBD VBN VBG', 'VBD VBN VBG RP'] then
44:             tense ← 'past_perfect_continuous'; voice ← 'active'
45:         else if POSsequence in ['VBZ', 'VBZ RP', 'VBP', 'VBP RP'] then
46:             tense ← 'present'; voice ← 'active'
47:         else if POSsequence in ['VBZ VBG', 'VBZ VBG RP', 'VBP VBG', 'VBP
    VBG RP'] then
48:             tense ← 'present_continuous'; voice ← 'active'
49:         else if POSsequence in ['VBZ VBN', 'VBZ VBN RP', 'VBP VBN', 'VBP
    VBN RP'] then
50:             tense ← 'present_perfect'; voice ← 'active'
51:         else if POSsequence in ['VBZ VBN VBG', 'VBZ VBN VBG RP', 'VBP
    VBN VBG', 'VBP VBN VBG RP'] then
52:             tense ← 'present_perfect_continuous'; voice ← 'active'
```

```
53:         else if POSsequence in ['MD VB', 'MD VB RP' ] then
54:             if VP[0][1] in FutureModals then
55:                 tense ← 'future'; voice ← 'active'
56:             else if VP[0][1] in ConditionalModals then
57:                 tense ← 'conditional'; voice ← 'active'
58:             end if
59:         else if POSsequence in ['MD VB VBG', 'MD VB VBG RP'] then
60:             if VP[0][1] in FutureModals then
61:                 tense ← 'future_continuous'; voice ← 'active'
62:             else if VP[0][1] in ConditionalModals then
63:                 tense ← 'conditional_continuous'; voice ← 'active'
64:             end if
65:         else if POSsequence in ['MD VB VBN', 'MD VB VBN RP'] then
66:             if VP[0][1] in FutureModals then
67:                 tense ← 'future_perfect'; voice ← 'active'
68:             else if VP[0][1] in ConditionalModals then
69:                 tense ← 'conditional_perfect'; voice ← 'active'
70:             end if
71:         else if POSsequence in ['MD VB VBN VBG','MD VB VBN VBG RP']
    then
72:             if VP[0][1] in FutureModals then
73:                 tense ← 'future_perfect_continuous'; voice ← 'active'
74:             else if VP[0][1] in ConditionalModals then
75:                 tense ← 'conditional_perfect_continuous'; voice ← 'active'
76:             end if
77:         else
78:             tense ← 'unknown'; voice ← 'unknown'
79:         end if
80:     end for
81: end procedure
```

Algorithm 3 Pseudocode for French tense and voice determination

```
1: procedure LABEL FRENCH VPS(List of VPs in a sentence)
2:     VPs ← liste of English VPs in a sentence
3:     MorphoTag ← concatenated morphological information of VP's constituting words
4:     Lemma ← list of lemmas of VP's constituting words
5:     Intransitives ← ['aller', 'arriver', 'décéder', 'devenir', 'échoir', 'entrer', 'mourir', 'naître', 'partir', 'rester', 'retourner', 'sortir', 'tomber', 'venir']
6:     for each English VP do
7:         if MorphoTag = 'V-subjonctif V-participepasse' then
8:             tense ← 'subjonctif'; voice ← 'passive'
9:         else if MorphoTag = 'V-indicatifpresent V-participepasse V-participepasse' then
10:            tense ← 'passé_composé'; voice ← 'passive'
11:        else if MorphoTag = 'V-indicatifimparfait V-participepasse V-participepasse' then
12:            tense ← 'plus_que_parfait'; voice ← 'passive'
13:        else if MorphoTag = 'V-indicatifpasse V-participepasse V-participepasse' then
14:            tense ← 'passé_antérieur'; voice ← 'passive'
15:        else if MorphoTag = 'V-indicatiffutur V-participepasse V-participepasse' then
16:            tense ← 'futur_antérieur'; voice ← 'passive'
17:        else if MorphoTag = 'V-indicatiffutur V-participepasse' then
18:            tense ← 'futur'; voice ← 'passive'
19:        else if MorphoTag = 'V-imperatifpresent V-participepasse' then
20:            tense ← 'impératif'; voice ← 'passive'
21:        else if MorphoTag = 'V-indicatifconditionnel V-participepasse' then
22:            tense ← 'conditionnel'; voice ← 'passive'
23:        else if MorphoTag = 'V-indicatifpresent V-infinitif V-participepasse' then
24:            if VP[0][5] in ['viens', 'vient', 'venons', 'venez', 'viennent'] then
25:                tense ← 'passé_ récent'; voice ← 'passive'
26:            else if VP[0][5] in ['vais', 'vas', 'va', 'allons', 'allez', 'vont'] then
27:                tense ← 'futur_proche'; voice ← 'passive'
28:            end if
29:        else if MorphoTag = 'V-indicatifpresent V-participepasse' then
30:            if lemma[0] = 'être' and lemma[1] in Intransitives then
31:                tense ← 'passé_ composé'; voice ← 'active'
32:            else
33:                tense ← 'present'; voice ← 'passive'
34:            end if
```

```
        else if MorphoTag = 'V-indicatifpasse V-participepasse' then
            if lemma[0] ='être' and lemma[1] in Intransitives then
                tense ← 'passé_antérieur'; voice ← 'active'
            else
                tense ← 'passé_simple'; voice ← 'passive'
            end if
        else if MorphoTag = 'V-indicatiffutur V-participepasse' then
            if lemma[0] = 'être' and lemma[1] in Intransitives then
                tense ← 'futur_antérieur'; voice ← 'active'
            else
                tense ← 'futur'; voice ← 'passive'
            end if
        else if MorphoTag = 'V-indicatifimparfait V-participepasse' then
            if lemma[0] = 'être' and lemma[1] in Intransitives then
                tense ← 'plus_que_parfait'; voice ← 'active'
            else
                tense ← 'imparfait'; voice ← 'passive'
            end if
        else if MorphoTag = 'V-indicatifpresent' then
            tense ← 'présent'; voice ← 'active'
        else if MorphoTag = 'V-indicatifimparfait' then
            tense ← 'imparfait'; voice ← 'active'
        else if MorphoTag = 'V-indicatifpasse' then
            tense ← 'passé_simple'; voice ← 'active'
        else if MorphoTag = 'V-subjonctif' then
            tense ← 'subjonctif'; voice ← 'active'
        else if MorphoTag = 'V-indicatiffutur' then
            tense ← 'futur'; voice ← 'active'
        else if MorphoTag = 'V-imperatifpresent' then
            tense ← 'impératif'; voice ← 'active'
        else if MorphoTag = 'V-indicatifconditionnel' then
            tense ← 'conditionnel'; voice ← 'active'
        else if MorphoTag = 'V-indicatifpresent V-infinitif' then
            if VP[0][5] in ['viens', 'vient', 'venons', 'venez', 'viennent'] then
                tense ← 'passé_ récent'; voice ← 'active'
            else if VP[0][5] in ['vais', 'vas', 'va', 'allons', 'allez', 'vont'] then
                tense ← 'futur_proche'; voice ← 'active'
            end if
        else
            tense ← 'unknown'; voice ← 'unknown'
        end if
    end for
end procedure
```