

Coreference resolution for spoken French

Loïc Grobol

▶ **To cite this version:**

Loïc Grobol. Coreference resolution for spoken French. Computation and Language [cs.CL]. Université Sorbonne Nouvelle - Paris 3, 2020. English. NNT : . tel-02928209v2

HAL Id: tel-02928209
https://hal.science/tel-02928209v2

Submitted on 9 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

École Doctorale 622 — Sciences du langage
Lattice, Inria

Thèse de doctorat en sciences du langage
de l'Université Sorbonne Nouvelle

Coreference resolution for spoken French

---

présentée et soutenue publiquement par

Loïc Grobol

le 15 juillet 2020

Sous la direction d'Isabelle Tellier[†] et de Frédéric Landragin

Et co-encadrée par Éric Villemonte de la Clergerie et Marco Dinarelli

Jury :
Massimo Poesio, Professor (Queen Mary University), Rapporteur,
Sophie Rosset, Directrice de recherche (LIMSI), Rapportrice,
Béatrice Daille, Professeur (Université de Nantes), Examinatrice,
Pascal Amsili, Professeur (Université Sorbonne Nouvelle), Examinateur,
Frédéric Landragin, Directeur de recherche (Lattice), Directeur,
Éric Villemonte de la Clergerie, Chargé de recherche (Inria), Co-encadrant,
Marco Dinarelli, Chargé de recherche (LIG), Co-encadrant.

Reconnaissance automatique de chaînes de coréférences en français parlé

Résumé

Une *chaîne de coréférences* est l'ensemble des expressions linguistiques — ou *mentions* — qui font référence à une même entité ou un même objet du discours. La tâche de reconnaissance des chaînes de coréférences consiste à détecter l'ensemble des mentions d'un document et à le partitionner en chaînes de coréférences. Ces chaînes jouent un rôle central dans la cohérence des documents et des interactions et leur identification est un enjeu important pour de nombreuses autres tâches en traitement automatique du langage, comme l'extraction d'informations ou la traduction automatique. Des systèmes automatiques de reconnaissance de chaînes de coréférence existent pour plusieurs langues, mais aucun pour le français ni pour une langue parlée.

Nous nous proposons dans cette thèse de combler ce manque par un système de reconnaissance automatique de chaînes de coréférences pour le français parlé. À cette fin, nous proposons un système utilisant des réseaux de neurones artificiels et ne nécessitant pas de ressources externes. Ce système est viable malgré le manque d'outils de prétraitements adaptés au français parlé et obtient des performances comparable à l'état de l'art. Nous proposons également des voies d'amélioration de ce système, en y introduisant des connaissances issues de ressources et d'outils conçus pour le français écrit. Enfin, nous proposons un nouveau format de représentation pour l'annotation des chaînes de coréférences dans des corpus de langues écrites et parlées et en nous en donnons un exemple en proposant une nouvelle version d'ANCOR — le premier corpus de français annoté en coréférence.

Mots-clés : anaphore, coréférence, réseaux de neurones artificiels, apprentissage artificiel, ressources annotées, corpus, formats d'annotation, traitement automatique du langage naturel

Coreference resolution for spoken French

Abstract

A *coreference chain* is the set of linguistic expressions —or *mentions*— that refer to the same entity or discourse object in a given document. *Coreference resolution* consists in detecting all the mentions in a document and partitioning their set into coreference chains. Coreference chains play a central role in the consistency of documents and interactions, and their identification has applications to many other fields in natural language processing that rely on an understanding of language, such as information extraction, question answering or machine translation. Natural language processing systems that perform this task exist for many languages, but none for French —which suffered until recently from a lack of suitable annotated resources— and none for spoken language.

In this thesis, we aim to fill this gap by designing a coreference resolution system for spoken French. To this end, we propose a knowledge-poor system based on an end-to-end neural network architecture, which obviates the need for the preprocessing pipelines common in existing systems, while maintaining performances comparable to the state-of-the art. We then propose extensions on that baseline, by augmenting our system with external knowledge obtained from resources and preprocessing tools designed for written French. Finally, we propose a new standard representation for coreference annotation in corpora of written and spoken languages, and demonstrate its use in a new version of ANCOR, the first coreference corpus of spoken French.

Keywords: anaphora, coreference, artificial neural networks, machine learning, annotated resources, corpus, annotations representation, natural language processing

Remerciements

Mes tous premiers remerciements vont à Isabelle Tellier, ma directrice de thèse, à qui je dois doublement mes débuts dans le monde de la recherche et de l'enseignement supérieur. Tu resteras pour moi un modèle et je ne pourrais jamais assez te remercier pour ta gentillesse, ton aide et ta confiance.

Je remercie de tout cœur Frédéric Landragin, mon directeur de thèse, qui a admirablement repris ce flambeau et dont l'infaillible optimisme et le soutien constant ont été les moteurs de toute la deuxième partie de cette thèse. Tu es incontestablement non seulement un excellent encadrant, mais aussi un collègue remarquable et un véritable ami. Je remercie aussi Marco Dinarelli, mon co-encadrant de thèse, à qui je dois mon introduction aux réseaux de neurones, plusieurs fructueuses collaborations et d'excellents moments, au Lattice comme en coreference. Je remercie encore Éric de la Clergerie, mon co-encadrant de thèse pour son soutien dans des moments compliqués, pour sa clairvoyances et pour nos très nombreuses discussions, j'ai grâce à toi de quoi m'occuper pendant plusieurs carrières et je ne sous-estimerai jamais les méthodes symboliques !

Je remercie les membres du jury d'avoir accepté d'évaluer cette thèse et sa soutenance dans des conditions loin d'être idéales, pour leurs nombreuses remarques, conseils et questions qui ont grandement contribué à l'amélioration et à la finalisation de ce travail. Des remerciements tout particuliers à Pascal Amsili, pour son aide dans la préparation de la soutenance à distance. Je remercie également les membres de mon comité de suivi de thèse, pour leur soutien et leurs conseils déterminants pour la fin de la thèse et pour la structure du présent manuscrit.

Je remercie le Lattice, ma *happy place* tout au long de cette thèse. Je pense tout particulièrement aux résident·e·s de Thésardland, à la solidarité sans égal, mais aussi à l'ensemble des membres du laboratoire, passé·e·s et présent·e·s, permanent·e·s et temporaires, sans qui ces années n'auraient pas été si douces.

Je remercie également l'ensemble des membres d'ALMAnaCH, pour leur accueil, leur soutien et l'ambiance stimulante de l'équipe qui m'a poussé tout au long de cette thèse à dépasser mes limites. Je remercie mes collègues enseignant·e·s du master pluriTAL, et en particulier Serge Fleury, Kim Gerdes, Sylvain Kahane, Jean-Michel Daube et Damien Nouvel, aux côtés desquels j'ai été heureux de pouvoir me consacrer à mon activité favorite, ainsi que l'ensemble de mes étudiant·e·s, et en particulier Natalia Kalshnikova et Veronika Solopova que j'ai eu le plaisir d'accompagner dans leurs travaux de recherche.

Je remercie les membres du consortium DEMOCRAT, pour leur implication dans un projet qui, s'il n'est finalement pas le cœur de cette thèse, a été déterminant pour mon travail durant ces années. Des remerciements particuliers à Serge Heiden pour ses encouragements, ses conseils et nos collaborations.

Je remercie les membres du groupe RITUEL, et en particulier Anne-Lyse Minard, Agatha Savary, Lotfi Abouda et la cabale de la coréférence : Emmanuel Schang, Adam Lion-Bouton, Sylvie Billot, Anaïs Lefeuvre-Halftermeyer et Jean-Yves Antoine, pour nos collaborations fructueuses — passées, présentes et futures ! — et pour leur amitié

Des remerciements tous particuliers à Iris Eshkol-Taravella, à qui je dois la découverte du TAL et mes premières expériences de recherche et avec qui j'ai toujours plaisir à travailler. Je remercie très chaleureusement Sophie Prévost et Benoît Crabbé, dont le soutien a été crucial pour toute la fin de cette thèse et avec qui je poursuis avec jubilation mes aventures de recherche en TAL.

Je remercie l'ensemble des *Cool Young Academics*, habituel·le·s et de passage, toutes et tous des *lovely humans*. En particulier, merci à Olga, Maximin, Sacha, Alice, Gael, Alexis, Ilaine, Zak, Marine C., Arthur B., KyungTae, Lara, Sanjay, Hicham, Élise, Pierre, Lucie, Ariane, Jade, Hyun Jung et Marie.

Je ne pourrais assez remercier le club des Chouettes Individus aux Nombreuses Qualités, Marie-Amélie, Marine, Auphélie, Mathilde, Miléna, Tian et Yoann ; vous êtes le houmous du houmous, du fond du cœur, merci. Mille mercis également à Mathieu, Mathilde, Paul, Émilie, Chloé, Florent, Adrian, Pedro, Clémentine, Alix, Jayjay, Sara, Arthur, Margot, Cassandra, Murielle, FX, Maxime, Nico, Manon, Lucas, Anaïs et toutes celles et tous ceux qui avant et pendant ces années m'ont soutenu sans faillir.

Enfin, je remercie ma famille et ma belle-famille, sur qui je sais pouvoir compter en toutes circonstances, et Hélène, spectatrice et actrice de cette aventure, partenaire de crime idéale et mon humaine préférée.

Contents

Front matter **ii**
Résumé . . . ii
Abstract . . . iii
Abstract . . . iii
Remerciements . . . iv

Contents **vi**

1 Introduction **1**
1.1 Coreference resolution . . . 1
1.2 Contributions . . . 4
1.2.1 A coreference resolution system for spoken French . . . 4
1.2.2 Leveraging imperfect knowledge . . . 5
1.2.3 A standard for multigenre coreference resources . . . 5
1.3 Outline . . . 5
1.4 Terminology and notations . . . 6

2 Coreference, anaphora and related phenomena **8**
2.1 Referring mechanisms in language . . . 8
2.1.1 Discourse models and coreference . . . 8
2.1.2 Referring mechanisms and interpretation . . . 10
2.2 Coreference in natural language processing . . . 13
2.3 Related tasks . . . 15
2.3.1 Named entity recognition and Entity linking . . . 15
2.3.2 Narrow coreference resolution . . . 17
2.3.3 Anaphora resolution . . . 19
2.4 Evaluating coreference resolution . . . 20
2.4.1 Mention detection . . . 20
2.4.2 Coreference resolution metrics . . . 21
2.4.3 Corpus-level evaluation . . . 26

3 Annotated resources for coreference **29**
3.1 Annotated resources for coreference . . . 29
3.1.1 Objectives . . . 30

3.1.2 Scopes . . . 31
3.1.3 Parallel and supporting annotations . . . 32
3.1.4 Ontonotes . . . 33
3.2 Resources for French . . . 35
3.2.1 Historical resources . . . 35
3.2.2 ANCOR . . . 36
3.2.3 DEMOCRAT . . . 40
3.3 Representing coreference annotation . . . 40
3.3.1 The URS metamodel for anaphora and coreference . . . 41
3.3.2 A TEI encoding and serialization for URS . . . 42
3.3.3 Application to ANCOR and DEMOCRAT . . . 44

4 Coreference resolution systems 49
4.1 Representations . . . 49
4.1.1 Mentions . . . 50
4.1.2 Mention pairs . . . 55
4.1.3 Mention sets . . . 56
4.2 Coreference resolution algorithms . . . 57
4.2.1 Ordering . . . 57
4.2.2 Link-centric algorithms . . . 59
4.2.3 Entity-mention models . . . 62
4.3 Mention detection . . . 64
4.4 Systems for French . . . 65

5 Automatic coreference resolution 67
5.1 System scope . . . 68
5.2 The DeCOFre system . . . 69
5.2.1 General architecture . . . 69
5.2.2 Representing text spans . . . 69
5.2.3 Mentions detection . . . 76
5.2.4 Coreference resolution . . . 76
5.3 Training . . . 79
5.3.1 Optimization algorithm . . . 80
5.3.2 Learning rate . . . 80
5.3.3 Regularization . . . 81
5.3.4 Mention detection . . . 83
5.3.5 Antecedent scoring . . . 84
5.3.6 Spans representations . . . 85
5.4 Experiments . . . 86
5.4.1 Hyperparameters . . . 87
5.4.2 Baseline results . . . 89
5.4.3 Transfer learning . . . 90
5.4.4 Hyperparameters influence . . . 91
5.5 Conclusion . . . 92

6 Explicit knowledge augmentations for coreference resolution systems 95
6.1 Raw features . . . 96
6.1.1 Span features . . . 96
6.1.2 Pair features . . . 97
6.2 Semi-supervised knowledge . . . 99
6.2.1 Static word embeddings . . . 99
6.2.2 Contextual word embeddings . . . 100
6.3 Shallow linguistic knowledge . . . 103
6.4 Structural linguistic knowledge . . . 104
6.4.1 Non-recursive structures . . . 105
6.4.2 Syntactic parsing . . . 106
6.4.3 Results . . . 107
6.5 Conclusion . . . 109

7 Conclusion and perspectives 111

Bibliography 113

List of Figures 150

List of Tables 151

Licence 153

Back 154

À Dino, mon grand-père, linguiste sans le savoir, à qui je dois ma rencontre avec l'études des langues. Et à Isabelle, grâce à qui je peux en faire mon métier.

Chapter 1

Introduction

1.1 Coreference resolution

This thesis is about coreference resolution for spoken French. A *coreference chain* is the set of linguistic expressions —or *mentions*— that refer to the same entity in a given document. *Coreference resolution* consists in detecting all the mentions in a document and partitioning their set into coreference chains. Coreference chains play a central role in the consistency of documents and interactions, and their identification has applications to many other fields in natural language processing that rely on an understanding of language, such as information extraction, question answering or machine translation.

(1) *[Susan Calvin]$_1$ had been born in [the year 1982], [they]$^\circ$ said, [which] made [her] seventy-five now. [Everyone]$^\flat$ knew [that]*. Appropriately enough, [U.S. Robot and Mechanical Men, Inc.]$_2$, was seventy-five also, since [it] had been in [the year of [[Dr. Calvin]$_1$'s birth]] that [Lawrence Robertson]$_3$ had first taken out [incorporation papers] for [what]$_2$ eventually became [the strangest industrial giant in [[man]'s history]]$_2$. Well, [everyone]$^\sharp$ knew [that], too.*

At [the age of [twenty]], [Susan Calvin]$_1$ had been part of [the particular Psycho-Math seminar at which [Dr. Alfred Lanning of [U.S. Robots]$_2$]$_5$ had demonstrated [the first mobile robot to be equipped with [a voice]]$_4$]$_6$. [It]$_4$ was [a large, clumsy unbeautiful robot]$_4$, smelling of [machine-oil] and destined for [the projected mines on [Mercury]].—But [it]$_4$ could speak and make sense.

[Susan]$_1$ said nothing at [that seminar]$_6$; took [no part] in [the hectic discussion period that followed]. [She]$_1$ was [a frosty girl], plain and colorless, [who]$_1$ protected [herself]$_1$ against a world [she]$_1$ disliked by [[a mask-like expression] and [a hypertrophy of intellect]]. But as [she]$_1$ watched and listened, [she]$_1$ felt [the stirrings of [a cold enthusiasm]].
(Asimov 1950)

Example 1 gives a possible output for a hypothetical coreference resolution system. Entity mentions are denoted by brackets and coreferent mentions are subscripted with the same index.

Even in such a short[1] example, we can already make some observations on the structure of the task. For instance, it seems that most mentions do not corefer with any other mention, and that, conversely, one entity —Dr. Susan Calvin— makes up most of the mentions, especially in the last paragraph. This suggests that the sizes of coreference chains are not uniformly distributed. We can also note that most mentions are noun phrases and pronouns, and that most pronouns corefer with a noun phrase that is not too far away from them.

This example also shows instances of many theoretical issues:

- Is [they]$^{\circ}$ really a mention of an entity, and if yes, what is the nature of this entity?
- What does [that]* refer to? The date of birth of Susan Calvin? Her age? Should we then also consider that the whole first sentence is also a mention coreferent with [that]*?
- Are the boundaries in [incorporation papers] correct for this mention or should they encompass the whole phrase 'incorporation papers for what eventually became the strangest industrial giant in man's history'?
- Are [everyone]$^{\flat}$ and [everyone]$^{\sharp}$ coreferent? What about the mentions of Dr. Calvin at different periods of her life?
- Is [a large, clumsy unbeautiful robot]$_4$ a mention of entity 4 or one of its attributes? Are coordinations like 'a mask-like expression and a hypertrophy of intellect'? Is [no part]?

From a technical point of view, this also shows that unlike many tasks in natural language processing, coreference resolution is a document understanding task that is hard to reduce to sentence-level processing or structural analysis. Indeed, understanding that [the strangest industrial giant in [[man]'s history]]$_2$ corefers with [U.S. Robots]$_2$ cannot be done reliably without building a model of the world described in this paragraph. It is also unlikely that an automatic system could rely on existing ontologies or knowledge bases to obtain a pre-existing model of this world, since it is entirely fictional.

Conversely, many references are relatively easy to resolve with simple heuristics: for instance, since [Susan] is the only noun phrase in the last paragraph whose grammatical gender is compatible with all the subsequent 'she', it would be relatively safe for an automatic system to bet on their coreference. Similarly, it is very unlikely that [U.S. Robot and Mechanical Men, Inc.]$_2$ and [U.S. Robots]$_2$ are not coreferent.

Coreference resolution —at least in limited forms— has a long history in natural language processing, both as a tool for other tasks, e.g. in Winograd (1972)'s language understanding system SHRDLU and as an issue in itself (Hobbs 1986). These early works were mostly focused on the specific case of pronominal coreference, traditionally formulated as the task of finding noun phrase antecedents for pronouns. The term of *coreference* itself rose to fame in natural language processing on the occasion of the MUC-6 (MUC Consortium 1995c) and MUC-7 (MUC Consortium 1998), where it was a part of a set of specific document understanding and information extraction tasks.

[1] Coreference resolution is usually formulated at the level of whole documents, in which case the example would be a whole book chapter.

During this time, many attempts at designing an accurate coreference resolution system were made, usually following the trends in other fields of natural language processing: moving from rule-based systems to data-driven machine learning systems and in the later years to the specific area of neural network-powered machine learning, led by the increasing availability of annotated datasets and computational power. These efforts mostly focused on coreference resolution for English and new paradigms often emerged from research on this language, encouraged by a greater availability of resources. However, the techniques developed for English have also been applied, usually successfully to other languages as soon as suitable corpora became available.

During these years, there has been a great diversity in approaches to coreference resolution, with none gaining a clear and durable upper-hand. This might be explained the difficulty of properly modelling a task that concerns many related but heterogeneous cognitive phenomena. This diversity manifests itself not only in the design of automatic systems, but also in the quantitative metrics used to assess their performances and in the frameworks and representations used in existing annotated data.

In this landscape, French appears as an odd language out: despite its usual status as a well-resourced language, no corpus with coreference annotation for French was available until 2014, far later than other well-resourced languages (such as Czech, Polish, German, Spanish or Catalan). This lack of resources was solved by the release in 2014 of ANCOR (Muzerelle et al. 2013a, 2014). However, while this corpus, comparable in size with the main coreference corpora for other languages, did provide the missing resource, it did so in another unusual way. Indeed, most coreference corpora are based on either standard written texts, or transcripts of planned speech and more rarely on spontaneous speech or non-standard writings, but ANCOR is made entirely of spontaneous speech with various degrees of interactivity. Therefore, research using ANCOR as a resource is confronted not only with the problem of coreference resolution in French but also of coreference resolution in spontaneous speech, both relatively unexplored tasks.

This is not to say that ANCOR is not a valuable resource. On the contrary: despite the lesser attention given to spontaneous speech in natural language processing, this genre has a lot of relevance for final applications —since spontaneous speech is the most common form of language produced and consumed by humans. It is also of a great interest for fundamental research: with a low availability of resources but close to a well-resourced language (in that case written French) it can serve as an experimental ground for work on truly lesser-resourced languages.

So far, however, no attempt at using ANCOR to design and evaluate a complete end-to-end coreference resolution system for French has been made public. Our objective in this thesis is to address this lack, by proposing such a system and exploring ways to overcome the issues caused by the specificities of ANCOR. These issues mostly derive from the differences between coreference in spontaneous speech and other genres: as suggested by works such as Antoine (2004) and Prince (1981), coreference chains do not follow the same patterns in spoken and written language, but these patterns have not been studied extensively enough to provide a simple way to transform a system design for written language into a system suitable for

spoken language. Furthermore, due to the lesser availability of resources for spoken languages, designing a coreference resolution system for spoken French using the same architecture as one designed for e.g. written English is not necessarily possible.

We propose to deal with these issues in the following manner:

- Since we lack a proper formalization of the specificities of coreference in spoken French, we do not attempt to design a system based on handcrafted rules, but instead defer the construction of these rules to data-driven machine learning techniques. Our concern is mostly to design suitable parametric language representations and processing architectures that can be optimized using the annotated data of ANCOR. Fortunately, the recent advances in machine learning gave us with suitable tools for this task and the current state-of-the-art in coreference resolution for other languages suggest that this is a viable approach with the amount of data available to us.
- Since resources for spoken French are scarce, we start by building a system that depends on as few resources as possible and ensure its viability before attempting to augment it with the imperfect knowledge that can be gathered from tools designed for written French. There again, recent advances in natural language processing brought by end-to-end neural architecture let us hope that this approach can indeed work in our situation.

1.2 Contributions

Our contributions in this thesis address several issues in coreference resolution for spoken French: designing a knowledge-poor coreference resolution system suitable for this language and genre, asserting the relevance of noisy knowledge sources for this task and improving the state of coreference resources for French by proposing a unified representation for coreference copora in both written and spoken language.

1.2.1 A coreference resolution system for spoken French

We propose DeCOFre, an end-to-end coreference resolution system, capable of identifying entity mentions in raw documents and specifically designed for transcriptions of spontaneous spoken French. This system does not rely on external knowledge[2] and its architecture should therefore be generalizable to other genres and languages as long as an annotated dataset of a size similar to ANCOR is available.

The architecture of this system is based on both components from neural coreference resolution systems for English (K. Lee et al. 2017, 2018) and general-purpose representation techniques coming from neural systems for other natural language processing tasks (Cross and Huang 2016; Devlin et al. 2018; Ling et al. 2015). Quantitatively speaking, the results obtained are better than those obtained by the only previous work on unrestricted coreference resolution for French (Désoyer et al. 2015a), which did not propose a way to detect entity mentions and used extensive gold-standard knowledge. This makes DeCOFre the first proper end-to-end

[2]Beyond the use of pretrained word embeddings, which we show in section 6.2 to be unnecessary although they provide some tangential benefits.

coreference resolution for French and the first coreference resolution system designed for and evaluated on spoken language transcriptions.

1.2.2 Leveraging imperfect knowledge

Having produced a knowledge-poor coreference resolution system, we propose several extensions of this system using knowledge sources of different types: from heuristics on the raw material of ANCOR that do not need external resources to knowledge extracted in semi-supervised ways from raw linguistic material, shallow linguistic processing and structured linguistic processing such as syntactic parsing. None of these knowledge sources were designed for spoken French, but either for heterogeneous web texts or standard written French.

We find that the relevance of these resources is highly variable, and in particular that general-purpose representations learned in semi-supervised ways are much easier to integrate in our system and improve it significantly, while knowledge relying on explicit linguistic processing does not seem to provide many improvements.

1.2.3 A standard for multigenre coreference resources

Finally, in order to support our usage of ANCOR in designing our system, we provide a new version of this corpus, including various improvements to its structure and content, correcting some mistakes and provide a reference word-level segmentation and standard split that we hope will help others to reproduce our experiments and compare our results to those of future works.

This resource comes in a new format, built upon the recommendation of the Text Encoding Initiative (TEI consortium 2020) in order to move from the ad-hoc representations of previous versions of ANCOR. This format is designed to allow coreference annotations for all documents encodable in TEI formats, including that of the DEMOCRAT corpus for coreference in written French (Landragin 2016) released during the elaboration of this thesis and to which we participated. It also allows for other types of annotations beyond coreference and can serve as a base in the conception of rich multi-annotated corpora using a large variety of materials.

1.3 Outline

This remaining of this document is organized as follows: in chapter 2, we give a discourse-oriented model of coreference along with the terminology that we adopt in this work, describe the task of coreference resolution as it is understood in natural language processing, outline some popular related tasks. We then provide definitions for the quantitative measurements used to assess the performances of coreference resolutions systems.

In chapter 3, we give an overview of the existing resources for coreference with elements of typology that we apply to the currently most popular resources for coreference resolution — the Ontonotes corpus (Hovy et al. 2006; Pradhan et al. 2007a) — and the existing resources for French. We also describe our contributions in the form of the XML-TEI-URS format for coreference annotations and its use ANCOR and DEMOCRAT.

In chapter 4, we describe the building blocks for existing coreference resolution systems in general and for French in particular, along with their respective requirements and the resources and algorithms that they rely on.

In chapter 5 we describe the knowledge-poor version of DeCOFre, our coreference resolution system in details, from its architecture to the training procedure that we used to optimize its performances of ANCOR and motivate some of our technical choices by empirical experiments.

Finally, in chapter 6, we describe our experiments of knowledge augmentations for DeCOFre, using publicly available resources —mainly designed from written French—, the results of these experiments and elements of interpretation.

1.4 Terminology and notations

Throughout this document, we use the following notations

Vector concatenation is noted by brackets. If $a = (a_1, ..., a_n)$ and $b = (b_1, ..., b_m)$ are vectors, $[a, b] = (a_1, ..., a_n, b_1, ..., b_m)$ is the concatenation of a and b.

Sequences are noted with indexed parentheses or in **bold** where parentheses would make a formula hard to read and the meaning is unambiguous. For instance, depending on the ambiguity of the context, the sequence $(h_1, ..., h_n)$ could also noted $(h_i)_{1 \leqslant i \leqslant n}$, $(h_i)_i$ or $\mathbf{h}$. If $S = (i_1, ..., i_n)$ is a sequence, we also write $(h_i)_{i \in S}$ for $(h_{i_1}, ..., h_{i_n})$.

Cardinality is noted with $|\cdot|$, for instance the cardinality of a set or an indexed family S would be noted $|S|$.

Neural network layers are noted as follows

Feedforward layers also called 'multilayer perceptrons' are noted FFNN. Formally a feedforward layer of depth $n \in \mathbb{N}$ and dimensions $(k_0, ..., k_n) \in \mathbb{N}^{n+1}$ is defined as

$$\left| \begin{array}{rccc} \text{FFNN} : & \mathbb{R}^{k_0} & \longrightarrow & \mathbb{R}^{k} \\ & x & \longmapsto & \text{FFNN}(x) = (f_n \circ ... \circ f_1)(x) \end{array} \right. \quad [1.1]$$

with for all $i \in [\![1, n]\!]$,

$$\left| \begin{array}{rccc} f_i : & \mathbb{R}^{k_{i-1}} & \longrightarrow & \mathbb{R}^{k_i} \\ & x & \longmapsto & f_i(x) = a_i(\mathrm{W}_i \times x + b_i) \end{array} \right. \quad [1.2]$$

where a_i is an elementwise non-linear function, W_i is a matrix of dimension k_i, k_{i-1} and $b_i \in \mathbb{R}^{k_i}$.

Long Short-Term Memory layers (Hochreiter and Schmidhuber 1997) are noted LSTM and are seen as operating on finite ordered sequences of vectors. More specifically, since we mostly use bidirectional LSTMs (Graves and Schmidhuber 2005; Schuster and Paliwal 1997), we note $\overrightarrow{\text{LSTM}}$ and $\overleftarrow{\text{LSTM}}$ their forward and backward directions and LSTM the stacked bidirectional version.

Gated Recurrent Units (Cho et al. 2014) are noted GRU. As with LSTMs, they are seen as operating on finite ordered sequences of vectors and we note $\overrightarrow{\text{GRU}}$, $\overleftarrow{\text{GRU}}$ and GRU their forward, backward and bidirectional versions respectively.

Definitions for technical terms can usually be found in either Jurafsky and J. H. Martin (2019) or (using their translation in French) in Cornuéjols and Miclet (2010). We defer to Diestel (2017) for graph theory terminology.

Chapter 2

Coreference, anaphora and related phenomena

2.1 Referring mechanisms in language

2.1.1 Discourse models and coreference

> *Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine will have to fulfill at least the following basic requirement. It has to be able to build a file that consists of the records of all the individuals, that is, events, objects, etc., mentioned in the text and, for each individual, record whatever is said about it.* (Karttunen 1976)

> *1. One objective of discourse is to enable a speaker to communicate to a listener a model s/he has of some situation. Thus, the ensuing discourse is, on one level, an attempt by the speaker to direct the listener in synthesizing a similar model.*
>
> *2. Such a discourse model can be viewed as a structured collection of entities, organized by the roles they fill with respect to one another, the relations they participate in, etc.* (Webber 1978)

> *Turning back to discourse, let us say that a **text** is a set of instructions from a speaker to a hearer on how to construct a particular **discourse-model**. The model will contain **discourse entities**, **attributes**, and **links** between entities. A discourse entity is a discourse-model object, akin to [Karttunen (1976)] **discourse referent** ; it may represent an individual (existent in the real world or not), a class of individuals, an exemplar, a substance, a concept, etc. Following Webber (1978), entities may be thought of as hooks on which to hang attributes.* (Prince 1981)

> *Language, then, is not merely interpreted with respect to worlds, models, contexts, situations, and so forth. Rather, it is involved in constructions of its own. It builds up mental spaces, relations between them, and relations between elements within*

> *them. To the extent that two of us build up similar space configurations from the same linguistic and pragmatic data, we may 'communicate'; communication is a possible corollary of the construction process.* (Fauconnier et al. 1994)

Following Fauconnier et al. (1994), Karttunen (1976), Prince (1981) and Webber (1978), for the purpose of this work, we focus on discourse (either uni- or multi-directional) as the construction of a model shared by all participants. This model is not simply a model of the real world, since the entities that it contains may be fictional, as the unicorn in example 2, or hypothetical, as the car in example 3.

(2) *Bill saw **a unicorn**. **The unicorn** had a golden mane.* (Karttunen 1976)

(3) *If Mary had **a car**, she would take me to work in **it**.* (Karttunen 1976)

It is not only a model of a physical world — real or fictional — either, since it can also involve facts (as 'le' in example 4) abstract notions or concepts ('the topological entropy of the tent function') and even nonsensical or paradoxical entities ('a four-sided triangle', 'a barber that shaves all those, and those only, who do not shave themselves' (Russell 1919) or 'the smallest uninteresting integer').

(4) *[...] faut que je le note d' ailleurs vous me **le** faites rappeler*

*([...] I must write it down by the way you remind me of **that**)*
(ANCOR, Muzerelle et al. 2014)

Finally, a discourse model might contain entities that depends on external factors, either because they are part of the general knowledge of a participant, as 'Noam Chomsky' in example 5, or because they refer to the situation at hand, as 'cette belle affiche' in example 6 (respectively *unused* and *situationally evoked* entities in the terminology of Prince (1981)).

(5) ***Noam Chomsky** went to Penn.* (Prince 1981)

(6) *Madame nous sommes autorisés par la municipalité pour mettre **cette belle affiche***

*(Madam we are allowed by the municipality to put **this beautiful poster**)*
(ANCOR, Muzerelle et al. 2014)

The entities and their attributes can change as the discourse unfolds, as in example 7, where it is clear that the 'zucchini' entity has significantly different attributes between the first and last sentences.

(7) *Pick **a ripe, plump zucchini**.[1] Prepare **it** for the oven, cut **it** into four pieces and roast **it** with thyme and smoked paprika for 1 hour. Serve **it** with white rice and enjoy **its** sweet and lightly spicy taste.*

In this framework, we hold it for given that entities are represented in language by parts of sentences (or utterances) — typically constituents or catena — the prototypical cases being noun phrases and pronouns. When this is the case, we will say that these parts are *mentions* and that they all *refer* to entities.

[1] With all due apologies to G. Brown and Yule (1983) for making their famous example more animal-friendly.

Given a document D and M, the set of all the mentions it contains, we will say that two mentions are *coreferent* or that they *corefer* if they refer to the same entity — which is clearly an equivalence relation on M.

Let $\bigsqcup_i C_i = M$ be the associated partition of M in equivalence classes, we will say that the C_i are the *coreference chains* of D (see section 2.2 for a motivation of the term *chain*). In other words, a coreference chain is the set of all the mentions that refer to a single entity.

We will also say that the first mention of an entity in a document is *discourse-new* and conversely that its other mentions are *discourse-old*. For convenience, in a slightly abusive borrowing from the terminology of anaphora, given a discourse-old mention *m*, we will say that the mentions preceding *m* in a discourse and coreferent with *m* are its *antecedents*.

Finally, we will say that a mention is *coreferring* if it is not the only element in its coreference chain. Conversely we will call[2] non-coreferring mentions *singleton mentions* or *singletons* when there is no ambiguity.

Note that this only define an *identity coreference* relation and does not address the case of *near-identity coreference* (Recasens et al. 2011a), which supersedes the notion of coreference, allows a much richer representations of the relations between mentions and entities and elegantly resolves a number of conundrums in coreference annotation. For the time being, the annotated resources are still too scarce to make near-identity identification systems practical — both in terms of conception and in terms of evaluation.

2.1.2 Referring mechanisms and interpretation

When referring to an entity (and possibly simultaneously introducing it in a discourse model), there are several possibilities for the emitter as to the choice of a referring expression. The primary constraint in this choice is that the chosen expression can reliably be interpreted by the receiver as referring to the correct entity —even if the choice can also be influenced by the need to convey attributes of this entity (such as in example 8) among other considerations.

(8) *Do you remember* ***Eyjafjallajökull****? The eruption of* ***this icelandic volcano*** *in 2010 stopped the air traffic for* 48 *hours.*

It is also to be expected that participants in an interaction try to conform to some sort of convention, e.g. along the lines of Grice's cooperation principles (Grice 1989), which usually implies that the chosen referring expression will be as concise as possible, all other things being equal.[3]

When referring to an entity, a speaker has a choice between several modes of reference. This choice is conditioned by the current *linguistic*, *situational* and *discursive* contexts. To be precise:[4]by *discursive context*, we mean the state of the discourse model ; by *situational context*,

[2]This identification between singleton mention and singleton entities has sometimes been the cause of heated debates —as have been other fine points regarding singletons.

[3]Let us stress, however, that this is a very rough approximation that should not be taken as an absolute formal requirement or — in the words of Ellen Prince — 'TO BE TAKEN WITH LARGE GRAIN OF SALT'.

[4]But by no mean exhaustive. For example this does not account for the fact that the extralinguistic context might

we mean the extralinguistic (including metalinguistic) elements of the interaction and their perceptions by the participants ; and by *linguistic context*, we mean the linguistic content of the interaction up to the production of a referring expression – most notably the order, recency and forms of preceding entity mentions.

This gives us a rough typology for referring expressions according to their respective degree of dependency to these contexts.

There are referring expressions whose interpretations depend essentially on encyclopedic knowledge that is assumed to be shared between discourse participants. *Proper nouns* (example 9) are the prototypical case for this mode of reference, although some *definite descriptions* (example 10) also fall in this category.

(9) ***Lovelace*** *was the only legitimate child of* ***poet Lord Byron*** *and* ***his wife Lady Byron***

(10) ***The first exoplanets to be discovered*** *are orbiting the pulsar PSR B1257+12, also known as 'Lich'.*

These expressions can be thought of as referring to a discourse entity *via* a link to a entity in that shared knowledge space.[5] As such, their interpretations does not rely on the pre-existence of the corresponding entities in the discourse model or the forms of their previous mentions. However, their forms are not completely independent of the context, since the assumptions made by the participants on their common knowledge have an influence on the degree of specificity that is used in these designations.

For instance, if a participant is assumed to ignorant of an entity, its first mention will also serve as an introduction. This can be seen in example 11: while the natures of 'Orléans' and 'Paris' can be assumed to be part of the general knowledge of fluent French speakers and are thus simply designated by their proper name. 'Lycée Pothier' is not as well know, yet it is not further specified. This is presumably because the speaker assumes that the interviewer has a some knowledge of the Orléans area —where the interview takes place and 'Lycée Pothier' is part of the general knowledge. Conversely, 'Argenteuil', a relatively well-known French city, is qualified as 'in the suburbs'. This might be due to the assumption from the speaker that their interlocutor —not being French themselves— might be unfamiliar with Argenteuil and that in this case this will give them enough information to infer the attributes of the 'Argenteuil' entity that are relevant at this point in the interaction: its nature and location.

(11) *[...] je suis né à* ***Orléans*** *[...] et j'ai fait mes études* ***au lycée Pothier*** *[...] j'ai travaillé à* ***Paris*** *également à* ***Argenteuil dans la banlieue*** *[...]*

([...] I was born in ***Orléans*** *[...] and I did my education at the* ***lycée Pothier*** *[...] I have worked in* ***Paris*** *also in* ***Argenteuil in the suburbs*** *[...])* (ANCOR, Muzerelle et al. 2014)

not be the same for all participants or intentional obfuscation from the emitter. Still, it should be sufficient for our needs in this work.

[5]For the purpose of this discussion, we will assume that encyclopedic knowledge functions as a background mental space specific to each individual and that participants in an interaction can make assumptions and deductions about one another's encyclopedic knowledge.

Note on the other hand that the actual referential nature of proper nouns is not always so clear-cut. For instance in example 12, the 'Nobel' part of the proper name 'Nobel Prize' does not necessarily create a discourse entity corresponding to Alfred Nobel, althoug in the terminology of Prince (1981), it does make it *inferable*, which allows the use of the pronoun 'he' in the next sentence.

(12) *Marie Skłodowska-Curie is the only person to have received the* ***Nobel*** *Prize in two different scientific fields. Had* ***he*** *known her, he would probably have agreed that this distinction was well-deserved.*

Similarly, the referential mechanism of *deictic expressions* mostly depends on a shared perception of the *situational context*. For instance, in example 13, the reference from 'ici' to the 'Orléans' entity is only inferable through knowledge of the location where the interaction is taking place.

(13) *[...] j'ai des camarades qui ne sont pas du tout d'Orléans et [...] qui ne trouvent pas la ville sympathique mais moi j'ai toujours vécu* ***ici*** *[...]*

([...] I have friends who are not from Orléans at all and [...] who do not think the city is nice at all but I, I have always lived ***here*** *[...])* (ANCOR, Muzerelle et al. 2014)

The situational context might include multimodal communication elements, as in example 14 where an accompanying gesture is the main factor of interpretation of 'celui-là'

(14) *j'aurais voulu la même chose en français mais pas* ***celui-là*** *[...]*

(I would like the same thing in French but not ***that one*** *[...])*
(ANCOR, Muzerelle et al. 2014)

Following Muzerelle et al. (2014), we also include discourse and metalinguistic deixis in this category, as 'la question' and '« on »' respectively in example 15.

(15) *nous avons dit «* ***on*** *» je vous quand je vous ai posé* ***la question*** *j'ai demandé est-ce qu'on fait assez pour les habitants d' Orléans ? et « on » ça représente qui pour vous alors ?*

(we said 'on'[6] *I you when I ask you* ***the question*** *I asked if 'on' did enough for the people of Orléans? and who is 'on' for you then?)* (ANCOR, Muzerelle et al. 2014)

Expressions that refer intensionally to entities via their attributes, their relations to other entities or both depend mostly on the discursive context. The precision of these intensional descriptions might range from complex predicates (example 16) to very general qualifications (example 17) and might be based on indirect information, such as morphological marks (example 18).

(16) *[...] alors comme vacances avec* ***mon amie de Tours*** *j'espère euh aller au à la Costa Brava [...]*

([...] so as vacations with ***my friend from Tours*** *I hope uhm to go to the the Costa Brava [...])* (ANCOR, Muzerelle et al. 2014)

(17) *I got on* ***a bus*** *yesterday and* ***the driver*** *was drunk.* (Prince 1981)

[6]Not translated to preserve the metalinguistic meaning in this example, but might be translated by **they** although **on** might also include first and second person referents, see Delaborde and Landragin (2019) for more details.

(18) *I saw a man with a dog yesterday and **it** was beautiful.*

Note that this mode of reference does not imply that the corresponding entity is already present in the discourse model at the time of utterance. Indeed, in example 17, 'a bus' and 'the driver' are actually the first mentions of their respective entities. In extreme cases such as example 19,[7] the intensional description can even be completely implicit.

(19) *She doesn't bike, though she owns **one**.*

Finally, the interpretation of some referring expressions mostly depends on the *linguistic* context. In these cases, the *discourse* reference to an entity might be *via* a linguistic reference to another expression. For instance, due to the syntax of dislocations in French, 'il' in example 20 is acting as a placeholder for the expression 'votre journal habituel la République du Centre', so they must refer to the same entity.

(20) *[...] votre journal habituel la République du Centre est-ce qu'**il** contient une rubrique sur le langage ?*

*([...] your usual newspaper La République du Centre does **it** include a column on language?)*
(ANCOR, Muzerelle et al. 2014)

Note that a given expression does not necessarily have a single mode of reference: for instance disambiguating between two entities with the same proper name might necessitate further qualifications when both are plausible in the current context.

2.2 Coreference in natural language processing

The term *coreference* in the context of natural language processing was introduced as part of the MUC-6 shared task (MUC Consortium 1995c) in the group of the 'SemEval' group of subtasks. Its definition was mostly informal and not linked to a particular theoretical framework of *reference*

> *The basic criterion for linking two markables is whether they are coreferential. Whether they refer to the same object, set, activity, etc. It is not a requirement that one of the markables is 'semantically dependent' on the other, or is an anaphoric phrase.*
> (MUC Consortium 1995b)

With 'markables' defined by

> *The coreference relation will be marked between elements of the following categories: **nouns**, **noun phrases**, and **pronouns**. Elements of these categories are **markables**. [...] The relation is marked only between pairs of elements both of which are markables.*
> (MUC Consortium 1995b)

and annotated as

```
<COREF ID="100">Lawson Mardon Group Ltd .</COREF>
said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF>
```

[7] Marcel Bollman is to thank for this one.

Therefore, in the context of this task, *coreference resolution* is defined as the process of detecting every mention m in a document that

- Is a constituent within a restricted set of types
- Is coreferent with at least one mention m' found earlier in the document, its 'antecedent'

and annotating both m, m' and the *directed* link $m' \leftarrow m$.

This formulation makes use of the fact that coreference is an *equivalence* relation to simplify the task. Consider the graph Γ over the set M of all mentions induced by the coreference relation: the coreference chains are its connected components, and since coreference is transitive, they are completely connected graphs — or *cliques*. Therefore, identifying coreference chains can be reduced to the identification one of its *spanning trees* and, since coreference is symmetric, it suffices to identify one of its spanning *directed trees*, *i.e.* to associate a unique antecedent to all but one of its mentions.

The search space is often restricted even further by constraining the direction of the links to be consistent with the order of appearance of the mentions in the discourse or in the extreme case, to allow only *chains*[8] (hence the expression *coreference chain*) or *stars*.[9] In the latter case, this formulation makes coreference resolution a proper generalization of the pronominal anaphora resolution task (see section 2.3.2) by identifying a canonical antecedent for every discourse-old mention. In this approach, the units of interest are the directed links between coreferent mentions and the entities are *latent*. This formulation is called the *mentions pair* or *link-centric* model of coreference (Recasens 2010).

It is also possible to use the dual, more discourse-oriented, approach of modelling *reference* directly and keeping *coreference* implicit. In that formulation it is not the coreference links between the mentions that are modelled but the links between the mentions and their corresponding entities in what is called the *entity-mention* model of coreference. It comes in two flavours: one (seen for instance in OntoNotes (Hovy et al. 2006)) models reference as a property of individual mentions, and deals with directed mention⟶entity links and makes the *mentions* the central units ; the other (seen for instance in ACE (Doddington et al. 2004)) models reference as a property of entities, deals with directed entity⟶mention links and makes *entities* the central objects.

Choosing among these three models of coreference has an influence in all the parts of the natural language processing treatment of coreference, which can be seen clearly in the diversity of the system evaluation metrics (section 2.4.2), annotation conventions (chapter 3) and system designs (chapter 4). Note however that these models are all equivalent and that the choice of one in a part of a task does not mandate the choice for the other parts, for example it is possible to evaluate an entity-centric entity⟶mention system (for instance CISTELL (Recasens 2010)) using a link-based metric (for instance MUC (Vilain et al. 1995)).

It is also possible to combine several approaches: e.g. a resource might include annotations

[8] Trees of maximal degree 2.
[9] Trees of diameter 2.

for both coreference chains and coreference links (as in the URS representation of ANCOR, see section 3.3). In practice, most implementations have in common a relative lack of details regaring the attributes of the entities and the type of (co)reference relations (but see section 2.3.3) and identify entities and their corresponding coreference chains.

As a result, in the context of natural language processing, given a document, *coreference resolution* is usually understood as the twofold task of

Mention detection Identifying the mentions in this document

Coreference resolution proper Identifying the coreference chains, with three alternatives

1. Associate every discourse-old mention with one of its antecedents (*mention-pairs* or *link-centric* model)
2. Affect every mention to the corresponding entity (*mention-centric entity-mention* model)
3. Partition the set of all mentions (*entity-centric entity-mention* model)

Note that these subtasks are not necessarily treated sequentially (see chapter 4).

A recurring question in practice is the treatment of singletons. Some implementations (such as MUC and OntoNotes) consider that since singleton mentions are not actually **co**referring, they should not be counted as mentions for a coreference task. As a result, for these shared tasks, the associated resources and the systems designed for them, 'mention' does not include singletons. The opposite point of view is that singleton mentions are still *referring* and that as such they still define coreference chains, if degenerated ones.

While both alternatives have upsides and wonsides, this choice clearly has a substantial impact on resource constitution, system design and evaluation and should be made explicit when describing a task, a resource or a system.

2.3 Related tasks

2.3.1 Named entity recognition and Entity linking

Named entities recognition is closely linked to coreference resolution since their common inception as subtasks in MUC-6. Ubiquitous in information extraction, the concept of *named entity* is notoriously hard to define formally, partly due to its origin in applicative rather than theoretical contexts. The most comprehensive attempt at a formal definition is perhaps due to Ehrmann (2008):

> *Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.*
>
> *(Given an applicative model and a corpus, we call* ***named entity*** *any linguistic expression that autonomously refers to a single entity of this model)* (Ehrmann 2008)

Thus, named entities *recognition* can be seen as a restriction of the mention detection subtask of coreference resolution, where the mentions that are considered are those that refer 'autonomously' —i.e. depending only on encyclopedic knowledge as defined in section 2.1.2— to a specific set of entities that are relevant for a specific applicative task. In practical implementations, the set of mentions is often further restricted to specific types of expressions. For instance, while arguably autonomously referring (Ehrmann 2008), definite descriptions and —more heterodoxically— speaker-referring pronouns, are seldom included in named entity recognition tasks.

This *detection* task is almost always associated with a task of *typological classification*, where detected mentions are sorted according to the nature of the corresponding entity in a given taxonomy. The extent and the granularity of these taxonomies is also task-dependent, from the relatively rough categories of MUC-6 (organisation, person, location, date, time, monetary amount, percentage) to the highly specific hierarchies of the Quaero corpus (Rosset et al. 2011) or Sekine and Nobata (2004) for instance.

Entity Linking is a further extension of named entity recognition, where the detection of named entities is associated with a *linking* task that associates the detected mentions with the corresponding entities, usually modelled as entries in a knowledge base. This can be seen as an extreme case of the entity type detection —where the types are of the maximum possible granularity— or as the analogue for named entities of the second subtask of coreference resolution. Historically, entity linking was initially proposed as an *extension* of coreference resolution to cross-document settings (Bagga and Baldwin 1998b) and by cross-pollination with word sense disambiguation challenges gradually evolved to its current form as large scale knowledge bases became available (Hachey et al. 2013).

While these two tasks are obviously closely related to coreference resolution, there has been relatively little synergy in system designs. This is in part due to the fact that their restricted set of types for mentions and entities allows for very convenient heuristics. In particular, most implementations only consider mentions that are either non-nestable (as in MUC-6 (MUC Consortium 1995a)) or whose structures are restricted to a limited degree of nesting (Rosset et al. 2011), contrasting with the potentially unrestricted nested structures of mentions in coreference resolution. With these restrictions, named entity recognition can be seen *segmentation* task, which in turn can be conveniently cast as a *sequence labelling* problem or —in the case of nested mentions— to successive sequence labellings (Dupont 2017; Straková et al. 2019). This formulation has the advantage of enabling the use of well-studied and efficient labelling algorithms. In particular, machine-learning algorithms such as linear-chain conditional random fields with Viterbi decoding (Lafferty et al. 2001) have yielded remarkable advances in named entity recognition in the last two decades.

As for entity linking, due to both historical reasons and the relative easiness of narrowing the referent of a mention down to a few homonymous entities in the reference knowledge base (Hachey et al. 2013), the techniques employed are closer to those used in word disambiguation and usually rely on ontological properties of the candidates entities. These approaches are not readily usable for coreference resolution, where entities are not necessarily well-defined and

their properties are not usually easily inferable (see section 2.1.1).

2.3.2 Narrow coreference resolution

In some contexts, it might be beneficial to attempt restricted versions of coreference resolution rather than the full version we describe in section 2.2. Indeed, since reference encompasses a great variety of mechanisms, it seems only natural that designing a system focusing on a narrower problem could be both easier and more useful in some situations. In particular, several tasks that focus only on certain types of mentions or certain types of relations have received significant attention, both historically before the advent of unrestricted coreference resolution tasks, and currently in the form of domain-specific tasks. We call these tasks *narrow* coreference resolution tasks and describe in this section some of the most important ones.

Note that coreference resolution as it is implemented in existing campaigns and datasets is itself actually narrow, as it usually focuses only on nominal and pronominal reference. The distinction we make is that in these cases, the narrowness is not usually a choice motivated by a theoretical or applicative objective, but rather a circumstantial necessity caused by the difficulty of actually annotating unrestricted coreference.

Pronominal reference resolution

Pronominal reference resolution is the restriction of coreference resolution to pronominal mentions, usually formulated as an antecedent-finding task. Formally: 'For every pronoun p in a document, find a noun phrase n such that p and n are coreferent'.

Historically, this task was the first type of coreference resolution implemented in natural language processing systems. Poesio et al. (2016c) identifies two reasons for this

> *There are two reasons for this focus on pronouns: a theoretical one – pronominal anaphora is much more governed by grammatical competence that full nominal anaphora – and a practical one – interpreting pronouns depends less on lexical, commonsense and encyclopedic knowledge than other types of anaphoric interpretation; hence, shallow approaches are more likely to achieve good results for this type of anaphora* (Poesio et al. 2016c, p. 78)

to which we add another practical reason: in downstream applications of coreference resolution, such as natural language understanding or machine translation, a reliable interpretation of pronominal reference is crucial, as a system can not rely solely on the content of a pronoun to derive a useful meaning.

This should not taken as meaning that pronominal reference resolution is a generally easier restriction of coreference resolution. If very simple baselines —such as choosing the closest noun phrase with compatible morphology (Recasens and Hovy 2010)— and naïve applications of government and binding heuristics can be sufficient for many pronouns, a non-negligible number of pronominal reference do not follow these criteria. To give but a single example Antoine (2004) reports up to 12 % of non-agreement between pronouns and their antecedents.

Moreover, ambiguous pronominal reference is both far from uncommon in attested corpora and challenging even for the current best systems (Webster et al. 2018). In fact, Levesque et al. (2012) propose to use 'Winograd schemas' —specific examples of pronominal reference— as an alternative to the Turing test. Specifically, considering the following pair of sentences:

(21) (a) *The trophy would not fit in the brown suitcase because **it** was too big.*

(b) *The trophy would not fit in the brown suitcase because **it** was too small.*
(Levesque et al. 2012)

reliably identifying the referent of 'it' in both sentence would involve a deep understanding of both the linguistic content of the sentences and the situation that they entail. It could actually be considered an AI-complete problem given certain guarantees to ensure that no side-channel is available. Studies on human performances on this task (Amsili and Seminck 2017) have shown that in the overwhelming majority of cases, human subjects do not experience any issue in solving Winograd schemas.

This is still an active area of research, with the recent work of Sakaguchi et al. (2019) providing more robust 'adversarial' Winograd schemas as an alternative to those introduced by Levesque et al. (2012) and those used as part of the GLUE (A. Wang et al. 2018) and SuperGLUE (A. Wang et al. 2019) benchmarks for natural language understanding.

Entity detection and tracking

Entity detection and tracking is an extension of the named entities recognition task to unrestricted mentions. Like named entity recognition, it focuses only on a few types of entities, but instead on focusing only on autonomously referring mentions, it consists in detecting *all* the mentions of these entities and their types (the 'detection' part) and linking coreferent mentions (the 'tracking' part). As such, unlike named entities recognition, it is a true narrow coreference resolution task.

In the context of the Automatic Content Extraction (ACE) shared task (Doddington et al. 2004) —the most famous implementation this task—, the types considered were 'Person', 'Organization', 'Location', F'acility', 'Weapon', 'Vehicle' and 'Geo-Political Entity', thus for instance 'a hoax' in 'President Bush says it's all a hoax' is not considered as a mention.

Event coreference resolution

Event coreference resolution is the restriction of coreference resolution to a specific set of entities that represent events of a limited number of types. It is usually treated separately from strictly (pro)nominal coreference resolution, mainly because the mentions that are considered are much more heterogeneous — ranging from pronouns to full propositions— and because their meaning is usually less readily accessible (Lu and Ng 2018). In addition to detection of event mentions and coreference resolution, implementations of this task also usually include the identification of some characteristics of events such as their types, participants, times or reality (Song et al. 2015).

While this narrow coreference resolution task is theoretically rather specific, neglecting it can have considerable consequences even on nominal coreference resolution. Consider for instance the case of example 22:

(22) *[...]* ***j'ai pas d'argent*** *euh je suis en dettes à l'EDF déjà de trois ou quatre millions enfin ça va* ***ça*** *ne durera pas de toutes façons* ***ça***

([...] ***I have no budget*** *uhm I have debts to EDF of already three or four millions, but it's alright,* ***it*** *won't last anyway* ***this****)* (ANCOR, Muzerelle et al. 2014)

In that case, not taking 'j'ai pas d'argent' into account would make the corresponding chain consist only of pronouns, which significantly lessens the usefulness of detecting this chain. Experimental results (H. Lee et al. 2012) have also shown that including event coreference as part of a coreference resolution system can significantly improve the quantitative results for both tasks. Accordingly, in some implementations —such as Ontonote's (Pradhan et al. 2007b)— event mentions are annotated if they corefer with at least one (pro)nominal mention, thus reaching a middle ground between purely (pro)nominal coreference annotation and the considerably more complex task of annotating all reference phenomena (see chapter 3).

2.3.3 Anaphora resolution

Following the definition of Poesio (2016), we say an expression is *anaphoric* if it '[depends] on the linguistic context, i.e., on objects explicitly mentioned or objects whose existence can be inferred from what has been said' (Poesio 2016, p. 24). As we have seen in section 2.1.2, the interpretation of referring expressions can be (and is in fact often) context-dependent, which makes these expressions anaphoric. But anaphoricity is not limited to the *identity reference* phenomenon on which we focus on in this work and also encompass the cases of near-identity coreference (Recasens et al. 2011a), discourse deixis (Webber 1991) — a superset of event coreference —, but also pro-forms, ellipsis and numerous others (see Hirst (1981) for a comprehensive review).

Anaphora resolution is the task of detecting anaphoric expressions and the contextual elements on which its interpretation depend, usually in the form of other linguistic expressions called their *antecedent*.[10] In practical implementations, it also usually includes resolving non-anaphoric identity references such as named entity coreference.

As we mentioned, the first implementations of a coreference resolution task in natural language processing was pronominal coreference resolution, which is usually anaphoric. As a result, and because of other uses of the terms *anaphora* and *coreference* in other areas of linguistics, a durable terminological conundrum has taken place, described in details by Poesio (2016, p. 38–39) where *anaphora* and *coreference* are used to denote various closely linked phenomena. This interpretation has also largely influenced research on coreference resolution, both in its terminology —for instance the use of the word *antecedent* for a preceding coreferent mention— and in its techniques —for instance the formulation of coreference resolution as an antecedent-finding task (see chapter 4).

[10]Note that in that case, the antecedent does not necessarily precede the mention, although it is by far the most common situation. The case where the mention precedes its antecedent is traditionally called 'cataphora'.

2.4 Evaluating coreference resolution

As we have seen in section 2.2, coreference resolution is a twofold task, with two very different components.

The first subtask —mention detection— has a strong syntactic component, and has historically been highly dependent on syntactic analysis. However, it is not a purely syntactic task, since not all noun phrases and pronouns are referential and reliably identifying these is far from trivial, even with expert linguistic analysis. For instance, discriminating non-referring, partially or fuzzily referring and fully referring pronouns is still an active area of research (Delaborde and Landragin 2019).

The second subtask —identifying coreference chains— is almost purely a discourse-level task and is usually more complex. Consequently, most of the focus in the literature has been on this subtask, both in terms of system conception and of evaluation design, with the result that this it is usually called 'corefrence resolution' itself and that the full task including mention detection is sometimes called 'end-to-end' coreference resolution. Another influential factor is the choice in the MUC-6 and MUC-7 shared tasks —the initial task codifiers— and in Ontonotes —the most common benchmark for coreference resolution in English— not to annotate (and therefore not to *evaluate* on) singleton entities. This conflation of mention detection and singleton detection makes it hard to disentangle the contribution of mention detection to the performance of end-to-end systems and provides little incentive to study and evaluate the performance on mention detection only.

Yet, there is a large body of evidence (Moosavi et al. 2019; Moosavi and Strube 2016; Popescu-Belis et al. 2004; Uryupina et al. 2016; Uryupina and Moschitti 2013) for the claim that the performance of mention detection has a crucial impact on the performances of end-to-end coreference resolution systems. Following Popescu-Belis et al. (2004), Recasens (2010) and Recasens and Hovy (2011), we thus treat mention detection as a task of its own, with dedicated evaluation.

2.4.1 Mention detection

Mention detection is in particular a *detection* task. As such, it is only natural to evaluate the output of a mention detection system —whether it comes from a dedicated module or is part of a holistic end-to-end processing— using the precision, recall metrics introduced by Kent et al. (1955) and the unweighted F-score defined by van Rijsbergen (1979) and popularized by Chinchor (1992) for information retrieval and commonly used for other detection tasks.

Formally, given a set of *t true samples*, i.e. elements to find in a collection, and a system output of *p positive samples*, i.e. elements —mentions in our case— detected by a system in the same collection, of which *tp* elements are in fact *true positive samples*, the *precision* P, the *recall* R and

the *F-score* F_1 are defined by

$$\begin{aligned} P &= \frac{tp}{p} \\ R &= \frac{tp}{t} \\ F_1 &= \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2 \cdot P \cdot R}{P+R} = \frac{2 \cdot tp}{t+p} \end{aligned} \quad [2.1]$$

Note that F_1 is in particular the unweighted *harmonic mean* of P and R, viz. the Pythagorean mean that is the most punitive to imbalance.

Yet this does not wholly solve the problem of mention detection evaluation. Indeed, if these metrics measure the global quality of a system output, one still has to define what it means to have successfully detected a given mention. According to the definitions given in section 2.1.1, a mention is a linguistic expression that refer to an entity. If for some mentions such as pronouns such as 'they' or proper nouns such as 'William Labov' the exact boundaries are clear, this leaves a wide range of options regarding what to include in a given mention.

Consider for example a text span such as 'Haden MacLellan PLC of Surrey, England' (Hirschman and Chinchor 1998). From a purely syntactic point of view, it behaves as a single element that could be replaced by a pronoun and that a detection system should therefore detect all of it. However, from a *task-oriented* perspective, detecting only 'Haden MacLellan' is not completely wrong, and it would seem reasonable to give partial credit to a system that would do so. This was dealt with in MUC by providing *maximal* and *minimal* spans for every mention, and report two separate sets of measures: one for strict detection of the maximal spans and one for any detecting any span that would be both a proper subspan of the maximal span and a proper superspan of the minimal span.

Yet, a double annotation for every mention is a time-consuming task and the time it takes is usually more useful to invest in other task for an annotation project. To address this problem, Recasens (2010) proposes an alternative 'lenient matching' where the syntactic head of a mention serves as its minimal span, and Moosavi et al. (2019) a more complex algorithm to infer minimal span from parse trees. However, both of these proposals still require access to a reliable syntactic analysis, which is not always the case, leaving open the question of how to evaluate partial mention detection in general. Options such as those sometimes used for named entity recognition of attributing partial credit for correctly detecting one of the two boundaries of a mention or using overlap ratios has not yet seen much interest for mention detection evaluation.

2.4.2 Coreference resolution metrics

Quantitative evaluation is a very important component in the definition of most natural language processing tasks and is usually conditioned on the definition of an evaluation *metric*, which evaluates the proximity of a system's output from a reference annotation. In that regard, coreference resolution stands out by its unusual number of competing metrics. Since the first proposition of Vilain et al. (1995) for a dedicated metric to evaluate and compare the

performances of coreference resolution systems (for the MUC-6 shared task), there has been at least five main counterproposals, with each one designed to overcome the shortcomings of the previous ones. In the current state of the art, there is no clear consensus (Poesio et al. 2016b) as to which of these metrics —which are not necessarily consistent with each other— should be used to rank coreference resolution systems.

There are multiple reasons for this situation: *historical* — the development and use of metrics being dependent on the availability of suitable resources— and *theoretical*, with the difficulty of formally defining a single degree of correctness for a given system output. Indeed, the addition of a single spurious mention in a chain can be a glaring error — for example a mistaken attribution in a Winograd Schema such as example 21 —, but grouping together nearly-coreferent mentions might be tolerable for most purposes —for example including 'le commun des soldats' in the emphasized chain in example 23.

(23) ***Les soldats qu'il avait commandés en Sicile*** *se donnaient un grand festin [...]* ***ils*** *se trouvaient nombreux,* ***ils*** *mangeaient et* ***ils*** *buvaient en pleine liberté [...]* ***le commun des soldats*** *était répandu sous les arbres [...]*

*(**The soldiers that he commanded in Sicily** were having a great feast [...]* ***they*** *were many,* ***they*** *were eating and* ***they*** *were drinking liberally [...]* ***The common soldiers*** *were spread out under the trees)* (Salammbô, Flaubert 1910)

Moreover, the human-judged severity of these errors is often inversely correlated with the propensity of existing systems to fall into them. This is the case for example 23: it is not terribly complicated for an automated system to learn that a partitive such as 'Le commun des' entails a restriction and prevents identity coreference, but as we saw in section 2.3.2, reliable automatic solving of Winograd schema is by design highly non-trivial and potentially AI-complete. This issue is also magnified by the relative scarcity of the phenomena causing these errors in actual corpora.

In the following, we consider a single document and the output of a coreference resolution system for this document and we note

- M_K, the set of all gold (or 'key' mentions), M_R the set of all system (or 'response') mentions and $M = M_K \cup M_R$ the set of all mentions
- K, the set of all gold coreference chains, which is a partition of M_K and R, the set of all system coreference chains, which is a partition of M_R

Except for MELA, these metrics all define a precision P and a recall R and a F_1-score defined as usual as the unweighted harmonic mean of P and R:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \qquad [2.2]$$

MUC

The MUC score (Vilain et al. 1995), was the first metric used for a full coreference resolution task. Vilain et al. (1995) defines the MUC precision and recall as

$$\begin{aligned} \mathrm{P} &= \frac{\sum_{r \in \mathrm{R}'}(|r| - |p(r, \mathrm{K})|)}{\sum_{r \in \mathrm{R}'}(|r| - 1)} \\ \mathrm{R} &= \frac{\sum_{k \in \mathrm{K}'}(|k| - |p(k, \mathrm{R})|)}{\sum_{k \in \mathrm{K}'}(|k| - 1)} \end{aligned} \qquad [2.3]$$

where $\mathrm{K}' = \{k \in \mathrm{K} \,|\, |k| > 1\}$ and $\mathrm{R}' = \{r \in \mathrm{R} \,|\, |r| > 1\}$ are the sets of non-singleton chains in key and response respectively and $p(x, \mathrm{E}) = x/\mathrm{E} = \{x \cap \mathrm{A} \,|\, \mathrm{A} \in \mathrm{E}\}$ is the partition of a chain x induced by a partition E of M.

These correspond to the precision and recall over the set of links between discourse-old mentions and their closest antecedent. Indeed, for a given non-singleton key entity k, there are $|k| - 1$ such links and $|p(k, \mathrm{R})| - 1$ of them are missing in the response. Thus, summing over K, R is the proportion of the key links that are present in the response. A similar analysis applies to P by switching the roles of the key and response.

Introduced as part of the MUC-6 conference, MUC is designed accordingly to its link-based coreference model and inherits its biases. The most notable of these biases is the choice not to annotate singleton mentions, reflected in the assumption in MUC that chains contain at least one link. This is not to say that singleton chains, in particular *response* singleton chains are completely ignored, since they still contribute to the $|p(r, \cdot)|$ term.

A more severe issue with the MUC score is that since it only consider links to the closest antecedent, mistakenly including a mention in an chain counts as an error in both precision and recall, while mistakenly merging two entities only counts as a single recall error (Bagga and Baldwin 1998a; Recasens and Hovy 2011). In fact, Finkel and Manning (2008) note that the edge heuristic of including all mentions in a single response chain would obtain a MUC F_1-score[11] of 88.2 %, which was higher than all the then-published coreference resolution system.

B^3

The B^3 score[12] (Bagga and Baldwin 1998a) was an answer to the aforementioned shortcomings of MUC. It is a mention-centric metric, originally defined[13] as

$$\begin{aligned} \mathrm{P} &= \frac{\sum_{m \in \mathrm{M_R}} \frac{|r(m) \cap k(m)|}{|r(m)|}}{|\mathrm{M_R}|} \\ \mathrm{R} &= \frac{\sum_{m \in \mathrm{M_K}} \frac{|r(m) \cap k(m)|}{|k(m)|}}{|\mathrm{M_K}|} \end{aligned} \qquad [2.4]$$

where $r(m)$ (resp. $k(r)$) is the response (resp. system) chain containing mention m.

[11] On the MUC-7 evaluation dataset.

[12] Or 'B-CUBED' in the original version, given without explanation for the name.

[13] In the unweighted version that is the most common in subsequent works. They also define a more general version with per-mention weights that amounts to computing a weighted average instead of an unweighted one.

This definition assumes that R and K are indeed partitions, which historically has not always been the case due to the presence of duplicated mentions in actual system responses. Due to the choice in the CoNLL campaigns not to enforce this condition, an alternative definition of B^3 was proposed by Pradhan et al. (2014):

$$\begin{aligned} P &= \frac{1}{|M_R|} \sum_{r \in R} \sum_{k \in K} \frac{|r \cap k|^2}{|r|} \\ R &= \frac{1}{|M_K|} \sum_{k \in K} \sum_{r \in R} \frac{|k \cap r|^2}{|k|} \end{aligned} \quad [2.5]$$

However, this still does not address all the issues that can arise with duplicated mentions. In fact, when dealing with duplicated system mentions, the recall is not necessarily smaller than 1 and can even be driven to an arbitrary high values.

B^3 takes into account the correct identification of singleton entities, which was one of the motivations to create it and use it over MUC. However, the reward for correctly identifying singletons might be too high given the large ratio of singleton in actual documents: Recasens and Hovy (2011) found a B^3 score of 59 % on ACE-2004 for an all-singleton baseline while noting that the singleton rate in ACE is actually lower than in other corpora.

CEAF

The Constrained Entity Aligned F-Measure (CEAF) if a family of metrics introduced by Luo (2005) and designed to adress the issues arising from using B^3 to evaluate the ouput of end-to-end systems, which usually contain spurious mentions and can potentially contain duplicates of gold mentions. Luo (2005) define it as

$$\begin{aligned} P &= \frac{\sum_{r \in R} \phi(A^{-1}(r), r)}{\sum_{r \in R} \phi(r, r)} \\ R &= \frac{\sum_{k \in K} \phi(k, A(k))}{\sum_{k \in K} \phi(k, k)} \end{aligned} \quad [2.6]$$

where $\phi : \mathscr{P}(M)^2 \longmapsto \mathbb{R}^+$ is a scoring function and A is a one-to-one mapping[14,15] from key chains to response chains that maximizes $\sum_{k \in K} \phi(k, A(k))$. Of course, the value of these metrics is highly dependent on the choice of ϕ, which should intuitively be a measurement of how similar two entities are, with the constraints that for all $(E, F) \in \mathscr{P}(M)^2$, $\phi(E, F) = \phi(F, E) \leqslant \phi(E, E) \neq 0$ and that if $E \cap F = \emptyset$, then $\phi(E, F) = 0$.

The two commonly used variants are the 'mention-based' $CEAF_m$, which uses ϕ_3, the cardinal of the intersection, for ϕ

$$\phi_3 : (k, r) \longmapsto |k \cap r| \quad [2.7]$$

and the 'entity-based', $CEAF_e$, which uses the Sørensen-Dice coefficient ϕ_4 (Dice 1945; Sørensen 1948)

$$\phi_4 : (k, r) \longmapsto \frac{2|k \cap r|}{|k| + |r|} \quad [2.8]$$

[14] If K and R are not of the same size, they might be augmented with copies of the empty set.

[15] One such mapping can be obtained efficiently using the Kuhn-Munkres algorithm (H. Kuhn 1955; Munkres 1957).

The qualifications of 'mention-based' (corresponding to the mention-centric entity-mention model of coreference) for CEAF_m comes from the fact that $\phi_3(e, e) = |e|$, which makes it the average over all mentions m of an attribution score of 1 if the key and response chains corresponding to m are aligned and 0 else. Similarly, CEAF_e can be seen as the average overlap between the aligned key and response chains, since for all e, $\phi_4(e, e) = 1$ and thus correspond to the entity-centric entity-mention model of coreference.

CEAF solves most of the issues identified for B^3 relative to the evaluation of system outputs with spurious and potentially duplicated mentions by enforcing a single alignment between key and response chains. However, it suffers from the same issues regarding singleton mentions: when singletons are in majority in K, they make up most of the score, leaving relatively little room for meaningful comparisons (Recasens and Hovy 2011). It also has a bias[16] against oversplitting chains: since the alignment is one-to-one, splitting a large key chain in two response chain of similar size results in a heavy loss for a mistake that is not necessarily critical.

MELA (CoNLL)

The Mention, Entity, and Link Average score (MELA), introduced by Denis and Baldridge (2009) combines the previous metrics in an attempt to compensate for their different biases. It is simply computed as the weighted average of the F_1 score of MUC (a Link-based metric), B^3 (a Mention-based metric) and a CEAF metric (an Entity-based[17] metric). Despite its mostly empirical motivation, it gained a lot of exposure following the use of the unweighted variant for the CoNLL-2011 and 2012 shared tasks on coreference resolution, where it provided a single score to rank the participating systems, leading to the name 'CoNLL score' that is now common in the literature. While it could be argued that by combining three differently biased metrics, this one might have even worse biases, in actual evaluation campaigns using the OntoNotes dataset, the rankings given by the CoNLL score were usually consistent with those given by each individual metric. An assessment of this finding is still to be done for other implementations of the task.

BLANC

The BiLateral Assessment of Noun-phrase Coreference (BLANC), introduced by Recasens and Hovy (2011) and extended to the case $M_K \neq M_R$ by Luo et al. (2014) uses a variation of the Rand index (Rand 1971) to measure the cluster similarity between K and R by focusing on the *links* between mentions.

A *coreference* (resp. *non-coreference*) link is a couple m, m' of mentions such that $m \neq m$ and m and m' are coreferent (resp. non-coreferent). Let C_K (resp. N_K) be the set of coreference (resp. non-coreference) links according to K and C_R (resp. N_R) the one according to R. As seen in table 2.1, the BLANC index is the arithmetic mean of the F_1-scores for the detection of coreference and non-coreference links.[18]

[16] Considered by its authors to be a feature of these metrics rather than a bug.

[17] Interestingly, Denis and Baldridge (2009) used CEAF_m, which is arguably more mention-centric, but the subsequent uses — starting with the CoNLL-2011 shared task — used CEAF_e instead.

[18] See Luo et al. (2014) and Recasens and Hovy (2011) for the edge cases where these definitions would imply

Table 2.1: Definition of the BLANC metric

	Coreference	Non-coreference	Average
Precision	$P_C = \frac{\lvert C_R \cap C_K \rvert}{\lvert C_R \rvert}$	$P_N = \frac{\lvert N_R \cap N_K \rvert}{\lvert N_R \rvert}$	$\text{BLANC}_P = \frac{P_C + P_N}{2}$
Recall	$R_C = \frac{\lvert C_K \cap C_R \rvert}{\lvert C_K \rvert}$	$R_N = \frac{\lvert N_K \cap N_R \rvert}{\lvert N_K \rvert}$	$\text{BLANC}_R = \frac{R_C + R_N}{2}$
F_1	$F_C = \frac{2 \cdot P_C \cdot R_C}{P_C + R_C}$	$F_N = \frac{2 \cdot P_N \cdot R_N}{P_N + R_N}$	$\boxed{\text{BLANC} = \frac{F_C + F_N}{2}}$

BLANC was defined in reaction to the issues of CEAF and B^3 with singleton entities, which was becoming more glaring as more corpora with singleton mentions (in particular AnCora (Taulé et al. 2008) and SemEval 2010 (Recasens et al. 2010)) were released. However, it is not often reported in works on coreference resolution for English, due to the lack of singleton mentions in the most common benchmark for English — Ontonotes — and the well-documented hegemony of English in natural language processing literature (Bender 2019). This situation is also self-reproducing, since new works on coreference resolution have to compare to previous works (which was one of the reason to report MUC in the CoNLL evaluation campaigns despite its well-known flaws) there is little incentive to compute and report more recent metrics.

Finally, there is a certain degree of confusion as to the usage of BLANC for system mentions introduced by Luo et al. (2014): in that case, contrarily to the claims of Moosavi and Strube (2016), links where one end is a spurious system mention are *not* counted as valid non-coreference links. Therefore, as with CEAF, the addition of spurious system mentions actively degrades the final precision score.

2.4.3 Corpus-level evaluation

A lot has been written about the aforementioned metrics, comparing their advantages and disadvantages. Yet, there is a distinct lack in this body of work: all the definitions that we have given concern the evaluation of the output of a coreference resolution system *on a single document*. However, in actual evaluation campaigns, the evaluation usually concern multi-document corpora. To our knowledge, a standard way to obtain corpus-level evaluations from document-level evaluations has never been specified for any of these metrics.

Thus, the *de facto* standard is the one set by the reference CoNLL scorer (Pradhan et al. 2014). As of version 8.0.1,[19] the scheme used to derive corpus-level evaluation is defined similarly to the usual method of computing the micro-averaged precision and recall in multi-class classification tasks.

dividing by zero.

[19] https://github.com/conll/reference-coreference-scorers/tree/v8.01

Remember that in classification tasks, the *precision* for a given class c is defined as

$$\mathrm{P}_c = \frac{|tp_c|}{|p_c|} \quad [2.9]$$

where tp_c is the set of *true positives* samples for c and p_c is the set of *positives* samples.

This can be seen as the average *correctness* of class attributions (1 if the class was correctly attributed, else 0) over all positive samples for class c. The *micro-averaged precision* over the set of all classes $\mathscr{C}$ is the average correctness computed over all samples regardless of their class.

$$\mathrm{P} = \frac{\left|\bigcup_{c\in\mathscr{C}} tp_c\right|}{\left|\bigcup_{c\in\mathscr{C}} p_c\right|} = \frac{\sum_{c\in\mathscr{C}}|tp_c|}{\sum_{c\in\mathscr{C}}|p_c|} \quad [2.10]$$

or equivalently as the average of the per-class precisions weighted by the numbers of positives samples per class

$$\mathrm{P} = \frac{\sum_{c\in\mathscr{C}}|p_c|\cdot \mathrm{P}_c}{\sum_{c\in\mathscr{C}}|p_c|} \quad [2.11]$$

Note that $c \in \mathscr{C}$ is in fact the total number of samples, so this is also equal to the micro-averaged *recall* or *accuracy*. This is readily interpretable as an empirical estimation of the probability that a given sample obtained by sampling uniformly in the dataset would be correctly classified.

Now, as we have seen in section 2.4.2, the existing metrics for coreference all[20] define a precision and a recall by averaging a measure of correctness over some units. These units are either links, mentions or entities, depending on the chosen model of coreference.

If we note E_d the set of the units of interest in a system response for a given document d and s a measure of correctness, the precision for this response would be:

$$\mathrm{P}_d = \frac{\sum_{e\in\mathrm{E}_d} s(e)}{|\mathrm{E}_d|} \quad [2.12]$$

Proceeding by analogy with eq. (2.10), it then seems natural to compute the average precision for a corpus $\mathscr{D}$ as

$$\mathrm{P} = \frac{\sum_{d\in\mathscr{D}}\sum_{e\in\mathrm{E}_d} s(e)}{\sum_{d\in\mathscr{D}}|\mathrm{E}_d|} \quad [2.13]$$

which is what the reference scorer implementation does, and is again equivalent to computing the average of the per-document precisions weighted by the size of their respective set of units of interest

$$\mathrm{P} = \frac{\sum_{d\in\mathscr{D}}|\mathrm{E}_d|\mathrm{P}_d}{\sum_{d\in\mathscr{D}}|\mathrm{E}_d|} \quad [2.14]$$

In practice, the units of interest for the previously defined metrics are

MUc links between discourse-old mentions and their closest antecedent

B³ mentions

[20] For this analysis, following the reference implementation, we consider the coreference and non-coreference metrics in BLANC as separate metrics and compute the final F_1 score using their corpus-level values.

CEAF_m mentions

CEAF_e entities

BLANC links (coreference and non-coreference)

There is an issue with this definition: even for classification tasks, using *micro-average* measures supposes that there is a reasonable balance across classes. Depending on the task, this can mean either that the classes' numbers of elements are close, or that the minority classes are not particularly relevant. If this assumption is not valid, a micro-averaged measure will be biased toward the majority class, thus blurring the value of this measure to estimate the actual effectiveness of a system. Similarly, for coreference, micro-averaged measures are only relevant if the *documents* are of similar sizes — where *size* is the number of units of interest. Otherwise, a micro-average is not a faithful account of system performances on smaller documents. Since coreference is a *document-level* phenomenon, artificially truncating or merging documents in order to achieve size similarity is also not feasible.

For the existing metrics, this issue has the most effect on BLANC: for a document with N mentions, the number of *coreference* links is roughly linearly dependent on N, but the number of *non-coreference* links scales *quadratically* with N. Therefore, in a heterogeneous corpus, the contribution of the smaller documents to BLANC is marginal when using micro-averages. Moreover, note that number of units of interest per document depends not only on the document size, but also of its genre, its topic…

However, taking the opposite option of using *macro-averages* i.e. unweighted averages of the per-document scores does not seem reasonable either. That choice would lead to an equal contribution of every document to the final score, which could lead the final evaluation to be overly optimistic, since coreference resolution is usually easier for shorter documents.

We propose to take a middle ground: averaging per-document scores, weighted by the number of mentions. While not necessarily ideal, this has the advantage of yielding the same results as micro-averages for B^3 and CEAF_m and comparable results for CEAF_e.

We implement this approach in *scorch*,[21] our **s**corer for **cor**eference **ch**ains, along with a complete reimplementation of MUC, B^3, CEAF and BLANC with a focus on straightforwardness, legibility and faithfulness to the most recent definitions of these metrics. Our implementation is continuously checked against the complete test suite distributed with the reference scorer, thus ensuring that the document-level[22] scores are exactly the same.

[21] https://github.com/LoicGrobol/scorch

[22] The reference scorer does not provide a corpus-level evaluation test suite.

Chapter 3

Annotated resources for coreference

In chapter 2, we have given a brief overview of identity reference phenomena in language and described coreference resolution as it is understood in natural language processing. Modern natural language processing is heavily reliant on annotated resources, which is not surprising: designing a system that simulates a human capacity is easier when examples of humans demonstrating said capacity are available. Coreference resolution is no different in that regard, and if early efforts such as those of Hobbs (1986) had to make do without such resources, all the significant system design efforts of the last two decades relied on annotated corpora.

In this chapter, we study how coreference is annotated in existing resources (section 3.1), with a specific focus on resources for French (section 3.2) and make a proposal for a new annotation model and an accompanying standard representation that can be used for coreference, designed to be able to take into account a large variety of source materials consistently (section 3.3).

3.1 Annotated resources for coreference

Leech (2005) defines *corpus annotation* as

> *[T]he practice of adding interpretative linguistic information to a corpus* (Leech 2005)

and later

> *[A]nnotation is a means to make a corpus much more useful — an enrichment of the original raw corpus. From this perspective, probably a majority view, adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes.* (Leech 2005)

Generally speaking, an annotated resource is a mean to enable further research on the annotated phenomena, both from corpus linguistics and natural language processing perspectives. More precisely, from a natural language processing perspective, an annotated resource is a tool to *develop* natural language processing systems through rule design or machine learning, a tool to *evaluate* the performances of such a system and to study its behaviour in itself, and a tool to

compare the performances and behaviours of different systems.

As such, it would be hard nowadays to imagine developing a coreference resolution system without having access to a corpus of a sufficient size and quality whith coreference annotations. Consequently, the design of such a corpus, its contents and the choices made during its production strongly condition the systems developed with (or for) it.

In this section, we provide a rough overview of the characteristics of existing resources and their disparities and a brief description of the Ontonotes/CoNLL-2011/CoNLL-2012 (Hovy et al. 2006) corpus, a resource whose influence on the development of coreference resolution systems has been critical since its publication and its use in the eponymous shared tasks (Pradhan et al. 2012, 2011) and which explains some characteristics of the systems that we describe in chapter 4.

3.1.1 Objectives

Following Fort et al. (2011) and Habert (2000), we consider that

> *Quelle que soit sa richesse, une annotation est cependant toujours orientée par une tâche, même si cela est implicite.*
>
> *(Whatever its richness, an annotation is always shaped by a task, even when this is implicit.)* (Habert 2000)

Accordingly, a corpus with coreference annotations is built toward one or several goals that shape all the steps of its constitution, from the choice of raw linguistic material and the theoretical framework underlying its annotations to the actual annotation process and its distribution and preservation —along with the material constraints within which these are done. As noted by Leech (2005), this does not preclude further reuse and applications beyond this original goal, but it does have an influence on them, with existing annotations becoming less useful the further away their uses stray from their original purpose.

The original motivations for existing annotated resources for coreference can be roughly sorted in two categories: those where the study of referential phenomena is seen as an end in itself, and those were it is a tool in broader studies in natural language processing, corpus linguistics[1] or digital humanities in general. Borrowing from the terminology of Fort et al. (2011), we will call these respectively *final* and *intermediary* applicative goals.

For example, among the best known coreference corpora for English, MUC (MUC Consortium 1995c, 1998) and ACE (Doddington et al. 2004) are typical of corpora with *intermediary* applicative goals. In both cases, their coreference annotations were part of larger efforts of semantic annotations meant as tools to design and evaluate natural language understanding systems This is clearly visible in their design choices: coreference was seen as complementary to named entity recognition, resulting in the lack of annotations of singleton entities in MUC (since those entities that were neither a named entity nor part of a larger coreference chain were considered of lesser importance to the message understanding task) and the restriction of the entities in

[1] 'La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution' (Rastier 2002)

ACE to named entities. The influence of the end goal is also visible in their choices of base materials: written news and military reports for MUC, and news from diverse sources (written, spoken, filmed) for ACE whose target was explicitly 'extraction of information from audio and image sources in addition to pure text' (Doddington et al. 2004).

On the other hand, the ARRAU corpus (Poesio and Artstein 2008; Uryupina et al. 2019), explicitly designed to study anaphora, both from a linguistic perspective and to help designing anaphora resolution systems is built from raw materials in various genres, includes both written and spoken language and annotations for a large subset of reference and anaphoric phenomena.

Of course, this should not be taken to mean that these categories do not intersect, since improving system design and evaluation should also be beneficial for downstream applications. For instance the motivations given for GAP (Webster et al. 2018), a corpus of ambiguous pronominal reference with enforced grammatical gender balance, were to provide both a benchmark for existing systems on this particular kind of reference *and* a resource to design coreference resolution systems that would be more useful for downstream applications e.g. machine translation and information retrieval.

3.1.2 Scopes

As we saw in section 2.3, in of natural language processing, *coreference resolution* is more a spectrum of closely related tasks rather than a strictly defined task. It is only natural, then, that the same diversity exists among the existing annotated resources. These different conceptions reflect in part the applicative goals for these resources (see section 3.1.1) but also the theoretical choices of their authors and the material (including human and financial) circumstances of their constructions. The breadth and depth of the annotations in a resource necessarily constrain its possible uses. Corpus linguistic studies and coreference resolution systems can only use the features that are provided and machine learning systems can only learn the phenomena that are annotated. However, all annotations come with a cost, and one that is particularly high for coreference, since it involves processing of complex phenomena at the level of whole documents. More precisely, in the grid of analysis proposed by Fort et al. (2012), even the subtask of resolving pronominal references to pre-detected mentions is already at maximal complexity for the 'context weight', 'ambiguity' and 'tagset dimension' criteria,[2] which puts it among the most complex of the common annotations tasks.

The most notable of the disparities observable in the existing corpora for coreference include:

- The breadth (see section 2.3.2) of the annotated referential phenomena (possibly beyond identity). For instance ANCOR (Muzerelle et al. 2014) includes annotations for identity and bridging anaphora, but only for strictly nominal mentions, while Ontonotes (Hovy et al. 2006) includes verb mentions for event coreference but does not include annotations for bridging anaphora.
- The inclusion or exclusion of singleton mentions. While for Baldwin et al. (1998) 'the

[2]This assumes that this subtask is formulated as a labelling task, which might not reflect the actual annotation process, but is a relatively straightforward way to obtain a lower bound for the complexity of annotating coreference.

decision to annotate singletons is a bit of a philosophical issue' and MUC (MUC Consortium 1995c) and Ontonotes (see section 3.1.4) among others famously do not include singleton annotations, Kummerfeld et al. (2011), Uryupina et al. (2019) and Uryupina and Moschitti (2013) remark that this makes the development and evaluation of mention detection systems more complex.

- The additional features annotated for mentions (which can include named entity type, definiteness, syntactic and morphosyntactic features…) These can be supporting annotations (see section 3.1.3) or only added for mentions, as is the case with ANCOR (see section 3.2.2)
- The model or models of coreference used in the annotations, which has an influence on the expressivity of the annotations for relations and entities and should be coherent with the planned downstream applications for intermediary applicative goals. For instance, for corpora designed with applications to EDT (see section 2.3.2) such as ACE (Doddington et al. 2004), using a mention-centric model is much more sensible than a link-centric model, since the definition of the task implies the existence of explicit (rather than latent) entities.
- A less studied choice is the categorization of certain anaphoric phenomena. For example, in ANCOR, references to the participants of an interaction are annotated as exophoric singletons rather than coreferent mentions. Another contentious case is that of appositives, annotated as identity reference in MUC and ACE —a choice famously criticized by van Deemter and Kibble (2000) on the grounds of formal interpretation—, annotated separately in Ontonotes and ignored in many other corpora such as ANCOR.

3.1.3 Parallel and supporting annotations

Beyond annotations directly related to coreference resolution, many resources annotated with coreference also include other kinds of annotations. Indeed, for annotators —both human and machines— these annotations can reduce the complexity of the task, sometimes significantly.

The annotations most commonly found alongside coreference are syntactic annotations, which make the mention detection subtask far easier if the mention candidates are restricted to syntactic constituents (or subtrees in dependency syntax analysis) since they reduce the task of findings mentions boundaries to an easier classification task. They can also help for annotations of minimal spans (see section 2.4.1), which can for instance be reduced to detecting a core constituent or a head word. Finally, until recently (see chapter 4), most automatic coreference resolution systems relied on morphosyntactic and syntactic features, both for mention detection and for coreference resolution, which makes gold annotation of these features desirable in the perspective of shared evaluation tasks, since it removes a bias in the comparisons.

For these reasons, many of the available corpora with coreference annotations also include some form of syntactic parsing and in fact, most of these were actually annotated in syntax before the coreference annotation process: Ontonotes (Hovy et al. 2006), AnCora (Taulé et al. 2008), TüBa-D/Z (Hinrichs et al. 2004) and the Prague Dependency Treebank (Nedoluzhko et al.

2016) are examples of coreference corpora that are full treebanks. For other corpora such as COREA (Hendrickx et al. 2013), the Polish Coreference Corpus (Ogrodniczuk et al. 2016) or EPEC-KORREF (Soraluze et al. 2012), the annotations are limited to morphosyntax or chunks.

In these cases, it is common to consider these annotations as lower layers, and to use them to support coreference annotations, rather than as parallel annotations. For instance, in AnCora, mentions are directly annotated by labelling constituents in the syntax trees. This ensures the consistency between the different annotations, but it also requires that the support annotations are either gold-standard before the coreference annotation procedure or an additional precorrection.

3.1.4 Ontonotes

The corpus known in part of the recent literature on coreference resolution systems as 'CoNLL-2012' or sometimes 'CoNLL' began its existence as Ontonotes (Hovy et al. 2006), a large (2.9 Mwords for the 5.0 release (Weischedel et al. 2013)), multilingual (English, Chinese and Arabic), multi-genres (it includes journalistic data, web texts and transcriptions of TV and radio broadcasts and telephonic conversations) and multi-annotation (Weischedel et al. (2013) mentions syntax, propositions, word senses, named entities and coreference).

The initial goals of Ontonotse was to be a resource with comprehensive *semantic* annotations for two types of applications:

- Machine learning systems for the natural language processing tasks corresponding to these annotations, by providing them with a large annotated dataset.
- Downstream natural language processing systems (the authors identified machine translation, question answering and summarization), to encourage a shift towards the use of rich inputs instead of 'impoverished text models like bags of words or n-grams' (Weischedel et al. 2013).

Ontonotes indeed saw some uses as a resource for other related tasks such as named entity recognition (Finkel and Manning (2009) is famous example) and narrow coreference resolution (e.g. for pronominal coreference as a tool for summarization by Gillick et al. (2009)), but its first systematic use for coreference resolution was for a subset of it (120 kwords from its English part) in the first shared task of the 2010 SemEval workshop (Recasens et al. 2010).

From then, it rose to omnipresence with its use as a benchmark in the coreference shared tasks of the CoNLL 2011[3] (Pradhan et al. 2011) and 2012 (Pradhan et al. 2012). These campaigns were (and are arguably still are) the largest-scale shared tasks for coreference resolution with over 23 participants for the 2011 edition and 16 for the 2012 edition. The combination of the size and richness of the dataset and the availability of results for many systems made it hard for subsequent efforts in automatic coreference resolution (at least for English) to avoid working on the corresponding datasets and comparing their performances to those of the participants to the shared task.

[3]Where only the English part was used.

As a result, the dataset thereafter often referred to as 'CoNLL' in coreference literature became a *de facto* standard benchmark for coreference and its idiosyncrasies became those of most of the coreference resolution systems developed for its three languages. This, combined with the well-documented tendency of part of the natural language processing community to focus exclusively on English (Bender 2019), also had the effect that most of the recent advances in coreference resolution algorithms were first developed with and for Ontonotes or Ontonotes-like data. Recent community efforts, such as the creation of workshop on Coreference Resolution Beyond Ontonotes (Ogrodniczuk and Ng 2016, 2017) attempt to remedy this state of affairs.

Generally speaking, the coreference annotations in Ontonotes can be seen as extensions of those in MUC and ACE. They follow a mention-centric entity-mention model where coreferent mentions are given a common `@ID` attribute. Their most notable specificities are the exclusion of some phenomena (see Poesio et al. (2016a) for a more comprehensive and detailed examination):

- Pleonastic and generic pronouns are not considered to be mentions
- Coreference between generic mentions is not annotated
- Copular, predicative and appositive constructions are not annotated as identity reference (and are ignored in the CoNLL shared tasks, which is not unusual in coreference corpora in general)

In contrast, it does include event coreference by marking coreferent verb mentions.

Finally, as we mentioned in section 2.2 and section 3.1.2, the most controversial (and possibly influential) choice in Ontonotes is the non-annotation of singleton mentions. As a matter of fact, this choice is never actually made explicit in the release publication (Pradhan et al. 2007b) or the dataset documentation (Weischedel et al. 2013), which merely notes that 'mentions of the same entity, concept or event are co-referenced as IDENT', which is to be understood as implying that *non-coreferring* ones —i.e. singletons— are not annotated at all. A plausible motivation for this convention would be that since the choice was made to annotate verb mentions, annotating singletons would have involved either annotating most verbs or having different conventions for noun and verb mentions.

Beyond the effects on system design mentioned in chapter 4, this choice also had an effect on the very vocabulary used in the literature on coreference resolution. Indeed, the documentation of Ontonotes and of the CoNLL shared tasks use the term *mention* to refer to *coreferring* mentions (for instance in the title of one of the secondary tasks 'Predicted plus gold mentions'). Following this terminology, some works on mention detection —for instance Kummerfeld et al. (2011)— include singleton detection (and removal), which implies document-level processing, since being a singleton is not an intrinsic property of a mention.

The portion of Ontonotes used in the SemEval corpus includes automatic singleton annotations, described by the authors as

> *[A]ll NPs and possessive determiners were annotated as singletons excluding those functioning as appositives or as premodifiers but for NPs in the possessive case. In coordinated NPs, single constituents as well as the entire NPs were considered to*

> *be mentions. There is no reliable heuristic to automatically detect English expletive pronouns, thus they were (although inaccurately) also annotated as singletons.* (Recasens et al. 2011b)

which goes to show that there is no easy way to reconstitute this information short of manual reannotation.

3.2 Resources for French

French is usually counted as a well-resourced language, both in terms of theoretical linguistics works and in terms of language resources.[4] However, for a long time, there was a distinct lack of annotated resources for coreference in French, compared to those available in other well-resourced languages (see for instance Poesio et al. (2016a) for a recent cross-language summary of these). There is no easy explanation for this situation, though Salmon-Alt et al. (2004) attribute it to the combination of the lack of interest until 2002 for generic resources on anaphora and the lack of freely available resources with pre-annotations (particularly syntactic annotations) to support coreference annotations after that —which is still true: the recent large-scale annotation projects for coreference in French did start from unannotated data.

The complementary releases of ANCOR (Muzerelle et al. 2013a, 2014) for spoken French and DEMOCRAT (Landragin 2016) for written French put an end to this situation: there are now two large-scale corpora with coreference annotations for French, made publicly available. This unlocked the development of machine-learning systems for coreference resolution for French, of which our own work, presented in chapter 5 and chapter 6, on coreference resolution for *spoken* French is an example.

In this section, we give a brief overview of the resources for coreference in French made available before 2013, then go on to describe ANCOR —our reference corpus— and DEMOCRAT.

3.2.1 Historical resources

The state of coreference resources prior to 2002 is described by Salmon-Alt (2002) as lacking in both quality and quantity:

> *les ressources françaises annotées en relations anaphoriques sont insuffisantes, tant au niveau quantitatif qu'au niveau qualitatif* (Salmon-Alt 2002)

We reproduce in table 3.1 their inventory of the existing resources at the time, excluding those that only concern non-identity anaphora.

Of these resources, only Popescu-Belis (1999) was both[5] available and general enough, and it was far too small for data-driven natural language processing. Salmon-Alt (2002) aimed to address this lack by providing a publicly available large-scale corpus with coreference annotations, but

[4] At the time of writing, it stood at the third place in ELRA's map of language resources (Calzolari et al. 2010).
[5] ARCADE (Tutin et al. 2000) was later publicly released, but remained too specific.

Table 3.1: French corpora with anaphora annotations prior 2002 (from Salmon-Alt 2002)

Reference	Words (#)	Mentions	Open access
Bruneseaux and Romary (1997)	30 k	Characters, places and items	yes
Popescu-Belis (1999)	10 k	NPs	yes
Clouzot et al. (2000)	95 k	3rd person pronouns	unclear
Tutin et al. (2000)	1 M	closed class anaphoric expressions	no
Trouilleux (2001)	45 k	3rd person pronouns	no

the project fell short of that objective.[6]

A later initiative was DéDé (Gardent and Manuélian 2005), a corpus of definite description with a fine-grained typology, but it is both too small (40 kwords) and too specific (since it does not include annotations of pronouns) for general coreference resolution.

3.2.2 ANCOR

ANCOR (Muzerelle et al. 2013a, 2014) is the first public large-scale corpus with unrestricted nominal coreference annotation for French. It started with a proof-of-concept study by Muzerelle et al. (2013b) that resulted in a smaller corpus named CO2.

The initial goals of CO2 were —in the short term— to provide a systematic study of anaphora in spoken French and to study the influence of homogeneity constraints such as identity of number or grammatical gender in spoken language. In the long term, the authors hoped that this resource could help designing and evaluating of named entity recognition and entity detection and tracking systems, improving the analysis of anaphora for human-computer communication and evaluate the performances for spoken language of anaphora resolution systems designed for written language.

The motivation for ANCOR was to go further in terms of corpus size: since the current state-of-the-art coreference resolution systems were (and still are) machine-learning systems, the development of such a system for French required more data than was available in CO2 or in the other corpora for French that were either non-public or too narrow (see section 3.2.1). It was also an opportunity to integrate more diverse language and to refine the annotation procedure, building on the experience of CO2. Generally speaking, ANCOR is a corpus with mostly *final* applicative goals, with some long-term intermediary applicative goals, which have had relatively little influence on its design.

The base material for ANCOR consists of samples from three corpora: ESLO (Enquête SocioLinguistique à Orléans, Eshkol-Taravella et al. 2011), and OTG (Office du Tourisme de Grenoble) and Accueil UBS (Université Bretagne-Sud), both part of the Parole Publique corpus (Nicolas

[6] It then enjoyed a relatively short second life as part of the FReeBank project Salmon-Alt et al. (2004), now only reachable through the Internet Archive Wayback Machine.

et al. 2002). ELSO is only partially included in ANCOR and the included portion consists of two subparts: ANCOR-CO2, used in the pilot study and ANCOR-ESLO. Table 3.2 gives the relative contributions of these parts.

Table 3.2: ANCOR subcorpora dimensions (from Muzerelle et al. 2014)

	Duration		**Words**	
Corpus	h	%	#	%
ESLO-ANCOR	25	81.97	417 k	85.45
ESLO-CO2	2.5	8.20	35 k	7.17
OTG	2	6.56	26 k	5.33
UBS	1	3.28	10 k	2.05
Total	30.5		488 k	

All of these corpora are of transcribed *spontaneous* spoken French, with slightly different genres and transcription conventions between the ESLO and Parole Publique parts.

The ESLO parts are both from the ESLO1 part of ESLO (Lonergan et al. 1974), transcriptions of recordings from 1968 to 1971 in Orléans, France. These recordings were originally with the two goals of constituting an educational resource for French as a Foreign Language students and teachers and a sociolinguistic corpus, the 'soundscape of a city' (Baude and Dugua 2011). The parts kept for ANCOR are interviews of citizens of Orléans. Their form resembles that of sociological interviews (Weber and Beaud 2003), with prompts in the form of open questions from the interviewer and answers of variable length from the interviewee. The topics range from the interviewee's uses of language to their day-to-day lives, with the intent of leading them to talk about their perceptions of their city and their language. Since the objective in this kind of interview is to elicit personal observations, sentiments and thoughts from the interviewee, the questions and possibly phatic feedbacks from the interviewers are usually kept as short as possible and the answers can be long monologues, even though some of them are simple yes/no answers. It also contains some extra-interview interactions, such as interview set-up and small talk between participants.

The Parole Publique parts are transcriptions of much more utilitarian speech. They record interactions in physical presence at Grenoble's tourist office in 2001 for OTG and telephonic conversations from the reception of the Université Bretagne-Sud in 2013 for Accueil UBS. Both consist of short interactions of no more than a few minutes, where one of the participants (caller or visitor) seeks an answer to a question or asks for a document. Most of the interactions consist of clarifications and negotiations, sometimes with a few extra comments or some small talk. Utterances are much shorter than in ESLO, and the interactions are much more interactive and noisy.

ANCOR is a corpus for *anaphora*, including bridging anaphora, but strictly restricted to nominal and pronominal expressions. Generally speaking, the authors strove for *consistency* over completeness. For instance the adverb *demain* ('tomorrow'), though considered to be a mention is never annotated. This was motivated by the need for a strict syntactic characterization of

mentions, annotation experiments in the pilot campaign having shown that broader and fuzzier definitions were causing the annotators troubles. Similarly, as we mentioned in section 3.1.2, mentions of the participants of the interactions were annotated as singletons rather than coreferent mentions, since the authors considered that their coreference was independent of the context, i.e. non-anaphoric and therefore not to be annotated. This last choice is somewhat surprising, given that while the same argument could be made for named entities, these are still annotated as coreferent.

In addition to mentions, expletive pronouns are also annotated, to help system designers 'car ceux-ci peuvent tromper les systèmes de résolution des anaphores'. The pronouns in the fixed idioms *s'**il** vous plaît* and ***il** y a*, however, are not.

The following annotations are also provided for mentions:

- Gender, number and part of speech
- Definiteness (indefinite, definite, demonstrative or expletive form)
- Inclusion or not in a prepositional phrase
- Named entity type (for named entities)
- "NEW" for the first mention of a coreference chain

The representation of reference follows a link-centric model as is common[7] for anaphora corpora, since it gives a natural common representation for all anaphoric phenomena. Specifically, the chosen antecedent for a mention is the first *nominal* mention of its chain (thus excluding pronominal antecedents, even in cases of pronominal cataphora —where a pronoun is the first mention—).

The anaphora relations are also typed, with the following hierarchy:

Identity anaphora are split between

- **Nominal** anaphora, when the mention is a noun phrase. It falls between two subtypes
  - **Direct** annotated `DIRECTE` when the mention and its antecedent have the same (nominal) head
  - **Indirect** annotated `INDIRECTE` otherwise
- **Pronominal** anaphora, annotated `ANAPHORE` when the mention is a pronoun.

Bridging anaphora are split between

- **Nominal** bridging anaphora annotated `ASSOC`
- **Pronominal** bridging anaphora, annotated `ASSOC_PR`

Apart from transcription information (which are not annotations *per se* but rather parts of the rich source material), these corpora had not been the object of any kind of annotation. Most notably, they included no canonical word segmentation and no syntactic analysis. In contrast

[7] But not the only solution: a mention-centric representation would also be suitable for instance.

with comparable resources such as Tüba-D/Z (Telljohann et al. 2006), the PCC (Ogrodniczuk et al. 2016) or Ontonotes (Pradhan et al. 2007b), this meant that the identification of mentions with their exact character-level boundaries had to be done from scratch by the annotators. To mitigate this difficulty, the annotation procedure was done in two distinct phases: one for mention identification and one for mention linking, with adjudication rounds by an expert annotator between and after them.

The nature of the source material (speech transcriptions) and the initial goals (studying anaphora in speech), a format that allowed for accurate representations was a necessity for ANCOR. Utterance boundaries, overlaps, speakers, disfluencies, non-verbal events… are all crucial parts of spoken language and destroying this information would have seriously impaired the potential subsequent studies. It was also crucial to preserve the synchronization with the original audio signal: the original corpora are nor word- or phoneme-aligned but there are sufficiently many timestamps to do it automatically in the current state of technology, which would allow studies on the relations between anaphora and prosody.

For convenience reasons, the choice of the original work was then to keep the source material in its original Transcriber XML-TRS format (Barras et al. 1998). However, this format is not meant for extensive annotations besides those necessary for transcription and was not natively supported by existing annotation software. As a result, the annotators simply worked on XML-TRS serializations in Glozz as if they were raw texts.

While this removed the need for file conversions or the adaptation of an annotation software to Transcriber, this approach has two critical issues. One is practical: since the annotation in Glozz are stand-off with character offsets, and since XML serialization offers no guarantee of character-level stability (i.e. a single XML document can have different characters stream representations), in order to annotate XML documents with Glozz, one has to link the annotations not only to a specific version of these documents, but to a specific serialization, which can be problematic. In addition to this, even if character-level stability could be guaranteed, it would still forbid any subsequent modification of the source material (corrections or metadata change) or at least making annotation transposition to these new version very complex.

The other issue is more fundamental: overlaying a new structure on an already structured document without linking the two structure makes it hard to enforce structural constraints, such as the one imposed by the ANCOR annotation guidelines that a mention cannot cross utterances boundaries.

A step in this direction was taken by Désoyer et al. (2015b), who converted the original stand-off annotations of Glozz to custom inline annotations. The result was a format that is not usable in either Glozz or Transcriber anymore but it at least ensured the consistency between the two structure by forcing them into a common XML tree and removed the brittle dependency on character positions. However, this format was completely ad-hoc, mostly undocumented and hard to use in practice, which prompted our own standardization effort (see section 3.3).

3.2.3 DEMOCRAT

Started in 2016, the Description et modélisation des chaînes de références (DEMOCRAT, Landragin et al. 2018) project aims at building a coreference corpus for written French. The first version of this corpus[8] has been released in June 2019 and as of 2020 the project is reaching its end.

The publication of ANCOR in 2013 endowed French with its first corpus with annotations for unrestricted nominal coreference, but the specific nature of its source material (transcriptions of spontaneous spoken language), as interesting as it is, made it an outlier compared to existing corpora for other languages. In particular, it made comparisons of the performances of coreference resolution systems quite complex, since most of these systems and particularly state-of-the art systems were designed for and using the English part of the CoNLL-2012 corpus (of which only a small part consists of spontaneous speech). Beyond the differences in annotation conventions, that could have been lessened, the differences between reference mechanisms in spoken and written languages (noted by Prince (1981) for instance) make the simultaneous change of these two parameters (language and genre) problematic. In that context, one of the motivation of DEMOCRAT is to fill this gap, by providing a corpus that would be complementary to ANCOR.

DEMOCRAT's source material is a compilation of libre texts in written French, from fiction and news texts in modern French and a smaller collection of historical texts.

Unlike ANCOR, the annotations in DEMOCRAT are strictly limited to coreference and do not include other anaphoric phenomena. It also lacks the richer features provided in ANCOR for mentions such as mentions types, named entities or definiteness. However, it does include an automatic tokenization, and part-of-speech and lemma for its tokens, provided by Treetagger (Schmid 1994).

The first release uses the XML-TEI TXM format (Heiden et al. 2010) with stand-off XML-TEI-URS (see section 3.3) annotations. The published material uses annotations for both entity-mention models: every mention is annotated with an identifier for the corresponding entity, and chains ar represented by URS schemes pointing to all of their mentions.

3.3 Representing coreference annotation

The earliest attempt at devising a standard for coreference annotation comes from Poesio et al. (1999) in the context of the MATE project, as a proposal for a scheme to be used in the MATE workbench (Isard et al. 2000) and later refined in Poesio (2004) for the GNOME corpus. It durably influenced subsequent coreference corpora, most notably ARRAU (Poesio and Artstein 2008; Uryupina et al. 2019), LiveMemories (Rodríguez et al. 2010–0021) and EPEC-KORREF (Soraluze et al. 2012). It also influenced part of the annotations for the AnCora (Taulé et al. 2008), Tüba-D/Z (Telljohann et al. 2006) or the Prague Dependency Treebank (Nedoluzhko et al. 2016).

However, it only provides facilities to annotate coreference *links* — which is only natural since it was designed for general anaphora annotation rather than coreference. This is an issue

[8] https://www.ortolang.fr/market/corpora/democrat

when using an entity-mention corerefence model and led to some fragmentation from projects that chose another model, and Ontonotes (Hovy et al. 2006) or the Polish Coreference Corpus (Ogrodniczuk et al. 2016) for instance do not use it.

Regarding resources for French, as we have seen in section 3.2, the two main corpora — ANCOR and DEMOCRAT — use different models of coreference, with ANCOR using a link-centric model (entity-level annotations are planned but have not been yet been produced) and DEMOCRAT using an entity-centric model. In order to build an homogenous resource for French combining these two corpora and enable a consistent access to their annotations without constraining one of them to the other's model (and thus losing information), we had to devise a new representation for coreference annotations. Moreover, the specific constraints of the rich source material of ANCOR (that encode speech transcriptions) and our plans to enrich these corpora with further annotation encouraged us to devise a format that would allow for as much interoperability as possible with encodings beyond coreference.

In the remaining of this section, we describe our proposal for a model and a representation format for coreference annotations. This proposal is implemented in the XML-TEI URS used for the first release of DEMOCRAT and a new release of ANCOR, first described in Grobol et al. (2017a), Grobol et al. (2018b) and Grobol et al. (2018a).

3.3.1 The URS metamodel for anaphora and coreference

URS (Units, Relations, Schemes) is an annotation *metamodel* introduced by Widlöcher (2008) in the context of discourse analysis, but designed to be general enough for a large class of linguistic annotations. It is based on the three eponymous classes of annotations

Units are parts of the source material. In the first implementation of URS in Glozz (Widlöcher and Mathet 2012), they were restricted to *contiguous* segments, but in principle, they can be any relevant part.

Relations are binary *links* between two annotations, modelling binary relations, either directed or undirected.

Schemes are *n*-ary relations between annotations, either homogenous or structured.

Note that 'annotation' in the definitions above refer to all three kinds of annotations and that e.g. a relation can link two other relations, a unit and a scheme or any other combination. All annotations can also bear more information in the form of attached feature structures, which are themselves non-relational, i.e. a feature can not refer to another annotation.

The first use of URS for coreference annotations is almost fortuitous: as we mentioned in section 3.2.2, the members of the CO2 and ANCOR (Muzerelle et al. 2014, 2013b) annotation projects chose to use Glozz mostly for its ergonomics and its integrated inter-annotator agreement computation module. In ANCOR, the annotations are anaphoric relations, and they are naturally represented as directed URS relations. Later on, for the MC4 project (Mélanie-Becquet and Landragin 2014) that served as pilot study for the DEMOCRAT annotation campaign, the designers of the Analec annotation platform (Landragin et al. 2012) chose to use URS as their

base metamodel to ensure the compatibility with Glozz. As a result, their annotations of coreference are also formulated in URS terms: in that case, the annotations are coreference *chains* as chains of mentions, represented as URS schemes.

Consequently, when we started to work on a representation format for DEMOCRAT, with the intention of being as compatible as possible with ANCOR and the MC4 pilot corpus, it appeared natural to merge these representations. We propose the following representation for anaphora and coreference:

Mentions are represented URS *units*: they are the portions of the source material that are the observable parts of anaphoric and referential phenomena. In most corpora they are contiguous spans for convenience but that is not a necessary constraint: for instance in the case of overlapping speech or with disfluencies, it would be perfectly reasonable to consider non-contiguous mentions.

Anaphoric relations including, but not limited to *identity* reference are URS *relations*. In most settings, it is natural to use *directed* relations, either between a mention and an antecedent —for link-centric models— or between a mention and an entity.

Entities represented as URS *schemes*, by identifying them with their coreference chains.

This representation has the major advantage of being able to reify —and consequently to *annotate*— the three base objects for coreference (mentions, links and entities). This comes at the price of either redundancy or irregularity: since there are several ways to encode identity reference (as relations or as schemes), corpus designers have to either encode it several times (with the need to enforce consistency) or to chose one of the possible representations (harming cross-corpora compatibility). Our recommendation is to encode it both ways: annotate links between mentions and their canonical antecedents and coreference chains. This allows a uniform treatement of all anaphoric phenomena and annotation of entity-level properties (such as properties of an entity) and link-level properties (e.g. the mechanism of reference, see section 2.1.2). The consistency between the annotations should be automatically enforced by annotation softwares, which are already necessary for coreference annotation, and thus should not be a burden for the annotators themselves.

Note that this model does not impose many hard constraints on actual realizations: different projects and different resources can still use annotations conventions that suit their specific needs. For instance, it does not impose any rule for the choice of a canonical antecedent, or a specific typology of anaphoric links. This also allows for more complex annotations, for instance near-identity relations as described by (Recasens et al. 2011a) are easily described as URS relations, either between a mention and an entity or between two entities.

3.3.2 A TEI encoding and serialization for URS

In section 3.3.1, we described a model for coreference annotations as an implementation of the URS metamodel. However, while this solves the issues of the annotation *semantic*, it does not provide a standard annotation *representation* or actual *serialization*. Indeed, URS itself is not necessarily bound to a representation format and its first implementations in Glozz (Widlöcher

and Mathet 2012) and Analec (Landragin et al. 2012) use very different representations.

Theoretically, it might not be a problem for interoperability: as long as conversion tools are able to transform third-party serializations into abstract representations, third-party softwares could each use their preferred native representation and still be able to use this common model. This is still a problem in practice: no matter the goodwill of the community, it is not reasonable to expect extensive support for every arbitrary native representation from third-party software designers. This issue is by no mean new —the need for interoperability of data is older than computer science itself— and the solution is well known: standardization.

For documents, in the context of digital humanities, the Text Encoding Initiative (TEI, TEI consortium 2020[9]) is one of the longest-standing and most successful efforts of standardization. Its guidelines for document encoding encompass the needs of most topics in digital humanities (which for the purpose of this discussion includes computational linguistics) in terms of expressivity and versatility. Of course, this sometimes comes at the expense of brevity and straightforwardness, but that can hardly be completely avoided —and in any case, it is not our main concern here.

Accordingly, in our effort to develop an *interoperable* and *versatile* format for coreference annotation, using a TEI representation seemed to be the best choice, with two main motivations. Firstly, a TEI-backed format benefits from the work done during the 30 years of development of the standard, including battle-tested representations and support from a large panel of software solutions. Even without explicit support for the precise semantic of the URS metamodel, features such as the validation of properties or basic representation facilities thus come at no cost. Secondly, the TEI already support a variety of rich input formats: for instance transcription of speech via the `spoken` module, which is of crucial interest for us given the nature of ANCOR.

We are not the first to make that choice either: as we mentioned earlier, TEI-compatible encodings of reference were already a proposal of (Bruneseaux and Romary 1997) and current reference corpora such as the Polish Coreference Corpus (Ogrodniczuk et al. 2016) use TEI encodings. However, these uses are all strictly designed for referential and anaphoric annotations, while we propose to represent the whole URS metamodel, which has the potential to represent many other types of annotations (see section 3.3.3 for a concrete example).

We propose to represent

Units as `<tei:span>`[10] elements, whose purpose is to 'associate[] an interpretative annotation directly with a span of text' (TEI consortium 2020), marking its nature by setting its `tei:type` attribute to `"unit"`.

Relations as `<tei:link>` elements, which 'define[] an association or hypertextual link among elements' (TEI consortium 2020), with a `tei:type` attribute set to `"relation"`. The linked annotations are given by the `tei:target` attribute as a space-separated couple of

[9]Due to the grassroot and evolutive nature of the TEI, chosing a single authority to cite is not a trivial task. Readers not familiar with the TEI might want to start with Burnard (2014) rather than diving head first in the technicalities of the guidelines.

[10]In all of this section, we refer to the TEI concepts as of the 4.0 version of the P5 guidelines (TEI consortium 2020)

`teidata.pointer`. If a directed relation is used, the convention should be that the first target is the source of the link and the second one is the destination.

Schemes as `<tei:link>` elements with a `tei:type` attribute set to `"schema"`.

Associating features with these elements can be done straightforwardly, through the integration of the SemAF standard (ISO 2017), implemented in the TEI in the form of feature `<tei:f>` and feature structures `<tei:fs>`. This representation, with its serialization in XML forms the XML-TEI URS format.

Additionally, we also recommend a stand-off integration of these elements to TEI documents, preferably through the use of the `<tei:standOff>` element proposed by Romary (2017), and officially made part of the TEI guidelins in version 4.0.0. This is motivated by the common need in real-world corpora for parallel annotations (see section 3.1.3), which are much easier to manage in separate files (or at least separate locations) rather than as inline elements interleaved in the source material. It is also far more convenient for dealing with concurrent annotations of the same phenomena —either by different annotators or as different versions of the same annotation— and complex annotation structure, such as partially overlapping units —for which the strictly hierarchical representation of annotations as inline XML is not sufficient.

Finally, among the three types of pointing mechanisms for stand-off annotations identified by Bański et al. (2016), we recommend to use references to a tokenization of the source materials (by representing tokens as `<tei:w>` elements for instance) if at all possible rather than character offset-based mechanisms (that make the manipulation of the XML serializations tedious and error-prone) or `<tei:anchor>` elements that require ad-hoc manipulation of the source material —negating most of the advantages of stand-off annotations. This is of course not an ideal choice, as it requires the existence of a pre-existing tokenization, but this is an assumption that is often already made by annotation softwares and it should not be too much of a burden in many cases, since current automatic tools can provide a good enough tokenization in most cases.

3.3.3 Application to ANCOR and DEMOCRAT

Our initial motivation by designing the XML-TEI URS format was to provide a single format for the two main coreference corpora for French: ANCOR and DEMOCRAT. As we mentioned in section 3.2.2, the annotation formats for previous versions of ANCOR were ad-hoc combinations of specialized formats for speech transcription and coreference annotations that were not easily reconcilable. Our solution was to develop a TEI-backed annotation format (section 3.3.2) that could be used to add coreference annotations to TEI representations of speech transcription. From there, using this annotation format for DEMOCRAT would be straightforward, since the source material for DEMOCRAT, written texts, were easy to represent in TEI.

Porting ANCOR to XML-TEI URS

The first step was to convert the inline annotated version of ANCOR (Désoyer et al. 2015b) and to port its speech transcription encoding to a TEI one. This already had several advantages, since the TEI offered richer representations than those available in the native Transcriber (Barras

et al. 1998) format originally used for ANCOR. For instance, utterances overlap are dealt with in Transcriber by creating special overlap utterances, leading to a situation where a single phatic 'hm hm' —which occurs very frequently in the ANCOR— can effectively split a long monologue in two as in example 24, where it even splits a mention ('Université d'Orléans-Tours') in two parts.

(24) *—[...] c'est une place cette place que j'ai maintenant qui est ingénieur gestionnaire des campus de l'université d'Orléans-*

— hm hm

— Tours parce qu'ils cherchaient quelqu'un de jeune dynamique [...]

(It is a position that I have now which is engineer intendant of the Université d'Orléans-Tours since they were looking for someone young dynamic)
(ANCOR, Muzerelle et al. 2014)

The original annotation in ANCOR used a complex encoding (linking the two parts of the mention in a URS scheme) that introduced heterogeneity of representation for the mentions. Encoded in TEI, this is more easily represented as one single utterance for each speaker (with additional marks to indicate that an overlap occurs at some point), making the split mention contiguous again.

Then, in order to make the support of stand-off annotations easier, we went on to tokenize the transcriptions. This was done automatically using UDPipe (Straka and Straková 2017) and a small set of post-correction rules to ensure that this tokenization was compatible with the existing mentions (the guidelines of ANCOR rule out sub-word mention boundaries and we stayed true to them). Listing 1 gives a simplified example of the result of this conversion process.

From there, converting the inline annotations to stand-off was a relatively straightforward matter of translating `<anchor>` elements to identifiers of the newly created `<tei:w>` elements. Listing 2 is a simplified example of an annotation block representing two coreferent mentions and their coreference chain (which also includes a third mention).

Syntactic analysis for ANCOR

A more detailed account of the work described in this part can be found in Grobol et al. (2018b).

As we mentioned in section 3.1.3, many coreference corpora include parallel syntactic annotations, which helps both natural language processing systems and corpus linguistics research. Accordingly, we wanted to explore these possibilities for ANCOR and integrate syntactic analysis in our own system (see section 6.4.2). Since human syntactic annotations were out of reach considering our resources, we chose to try to integrate automatic syntactic parsing as a substitute. The actual parsing was done using Dyalog-SRNN (Villemonte De La Clergerie et al. 2017) trained on the French part Universal Dependencies v2 (Nivre et al. 2016). Accordingly, the syntactic annotations that we obtained were in *dependency* syntax, which is somewhat unusual for a coreference corpus, but has the advantage of being supported by a large multilingual corpus (Universal Dependencies) and being better suited than constituency syntax for syntactic

```
<div type="section" xml:id="s2">
    <timeline>
        <when absolute="3.531" xml:id="t2.0"/>
        [...]
    </timeline>
    <u start="#t7.0" who="#spk2" xml:id="u7"
     end="#t7.19">
        [...]
        <w xml:id="u7-w76">au</w>
        <w xml:id="u7-w77">moment</w>
        <w xml:id="u7-w78">où</w>
        <w xml:id="u7-w79">je</w>
        <w xml:id="u7-w80">me</w>
        <w xml:id="u7-w81">suis</w>
        <w xml:id="u7-w82">marié</w>
        <w xml:id="u7-w83">en</w>
        <w xml:id="u7-w84">juillet</w>
        <w xml:id="u7-w85">soixante-sept</w>
    </u>
</div>
```

Listing 1: Speech transcription annotations in ANCOR-AS

annotation of *spoken* French (Lacheret et al. 2014). Figure 3.1 gives an example of the syntactic analysis of part of an utterance in ANCOR.

Dependency syntax translates straightforwardly in URS terms:

Syntactic words can be represented as URS units, which also allows grouping several orthographic tokens in a single syntactic word

Dependencies can be represented as URS directed relations

The only modification to our previous formulation of XML-TEI URS was the addition of `<tei:expan>` elements to represent the expansion of contractions such as ‘du’ that are single orthographic tokens into several syntactic words (here ‘de le’).Listing 3 gives an example of the encoding of the dependency tree from fig. 3.1 in XML-TEI URS.

URS annotations for DEMOCRAT

Having proof-tested XML-TEI URS annotations in ANCOR, we went on to propose it for the release version of DEMOCRAT. This was done jointly with the integration of a URS annotation plugin in the TXM platform (Heiden et al. 2010). This integration was relatively easy, since the internal format of TXM was already built on TEI and it helped to develop its support for

```
<standOff>
    <annotation type="coreference">
        <spanGrp type="unit" subtype="mention">
            <span from="#u7-w76" to="#u7-w77" xml:id="m31"/>
            <span from="#u7-w84" to="#u7-w85" xml:id="m32"/>
        </spanGrp>
        <linkGrp type="relation" subtype="coreference">
            <link target="#m31 #m32" xml:id="r20"/>
        </linkGrp>
        <linkGrp type="schema" subtype="chain">
            <link target="#m31 #m32 #m40" xml:id="c12"/>
        </linkGrp>
    </annotation>
</standOff>
```

Listing 2: Coreference annotations in ANCOR-AS (fragments)

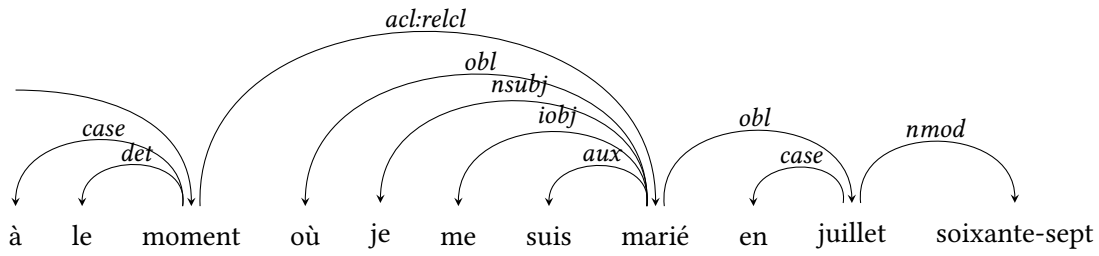


Figure 3.1: Syntactic analysis (subtree) for "*au moment où je me suis marié en juillet soixante-sept*"

parallel and versioned annotations. Conversely, the decoupling between annotations and source materials also allows a limited support for source versioning, since as long as the pointers supporting the annotations stay valid, it is not necessary to change the annotation.

Finally, by providing a single abstraction for a large class of annotations, this format also allows a uniform access to annotation in the form of the URS Query Language (URSQL) available in TXM since version 0.9 (Heiden 2019), which builds upon the concept of the GlozzQL (Mathet and Widlöcher 2011) language used in Glozz and allows complex combinations with the other specific querying languages available in TXM such as CQL (Evert and Hardie 2011), providing powerful interface for corpus linguistics inquiries (Quignard et al. 2018).

```
<standOff>
    <annotation type="syntax">
        <div type="tree" xml:id="tree10">
            <div type="multiword-token">
                <expan xml:id="tree10-w6-7" n="6-7" corresp="#u7-w76">
                    <w xml:id="u7-w76.1">à</w>
                    <w xml:id="u7-w76.2">le</w>
                </expan>
            </div>
            <spanGrp type="unit" subtype="word">
                [...]
                <span target="#u7-w76.1" n="6" xml:id="tree10-w6" ana="#tree10-w6-fs"/>
                <span target="#u7-w76.2" n="7" xml:id="tree10-w7" ana="#tree10-w6-fs"/>
                <span target="#u7-w77" n="8" xml:id="tree10-w8" ana="#tree10-w6-fs"/>
                [...]
            </spanGrp>
            <linkGrp type="relation" subtype="dependency">
                [...]
                <link target="#tree10-w8 #tree-10-w6" xml:id="tree10-d6" ana="#tree10-d6-fs"/>
                <link target="#tree10-w8 #tree-10-w7" xml:id="tree10-d7" ana="#tree10-d7-fs"/>
                <link target="#tree10-w3 #tree-10-w8" xml:id="tree10-d8" ana="#tree10-d8-fs"/>
                [...]
            </linkGrp>
            <div type="dependency-fs">
                [...]
                <fs xml:id="tree10-d7-fs">
                    <f name="type"><symbol value="det"></f>
                </fs>
                <fs xml:id="tree10-d8-fs">
                    <f name="type"><symbol value="obl"></f>
                </fs>
                [...]
            </div>
            <div type="word-fs">
                [...]
                <fs xml:id="tree10-w6-fs">
                    <f name="upostag"><symbol value="DET"></f>
                    <f name="Definite"><symbol value="Def"></f>
                    [...]
                </fs>
                [...]
            <div>
        </div>
    </annotation>
</standOff>
```

Listing 3: Syntactic annotations in ANCOR-AS corresponding to fig. 3.1

Chapter 4

Coreference resolution systems

In this chapter, we review the existing techniques, architectures and algorithms used in existing coreference resolution systems.

In section 4.1, we describe the input representations strategies for the three types of objects at the core of coreference resolution: mentions, mention pairs and mention sets.

Section 4.2 gives an overview of existing coreference resolution algorithms, sorted by their model of coreference (mention-pair and entity-mention) and their document processing strategies (discourse-ordered and global).

We will treat mention detection separately in section 4.3, since depending on the understanding of 'mention' in tasks definitions (see section 2.2) and resources (see section 3.1.2) it can be either a mostly syntactic task or so tightly tied to coreference resolution that it is a mere by-product of the coreference resolution algorithms used by end-to-end systems.

Finally, in section 4.4 gives a brief summary of the few existing works on coreference resolution for French.

4.1 Representations

A fundamental problem in natural language processing representing objects and concepts that are not always easy to define formally in logical or numerical forms that a machine is able to process. Even assuming that we are only concerned with unstructured text, it is possible to simply represent the inputs to natural language processing systems as streams of characters, but that would make their design overly complex for all but the simplest tasks. Consequently, the representations used in natural language processing are usually based on higher levels representations informed by linguistic theories —words, sentences…—, themselves usually represented as sets of *features*: symbolic or numerical properties that designed to be both sufficiently expressive and relevant for a given task.

Historically, features were defined directly by experts alongside the rules of the systems that used them. With the increasing availability of processing power and linguistic data (both raw

and annotated), the design of both[1] features and rules tends to be delegated to machine learning algorithms, and the role of system designers has moved towards the design of machine learning architectures. This is not to say that expert knowledge does not play a role in this design process: on the contrary, it is critical to the design of architectures that are suitable interfaces between the linguistic data to be processed and its use in the subsequent parts of these systems. It is also currently irreplaceable in the quasi-craftsmanship process of training these systems.

For coreference resolution, the unavoidable objects are *mentions*, but most algorithms also use higher level representations of either mentions *pairs* or mentions *sets*, including their final targets: coreference chains — usually conflated with entities as is common in modelling coreference resolution for natural language processing. Some algorithms also call for representations of composite objects such as cluster-mention or cluster-cluster pairs. The representation strategies for these objects have mostly followed the same trends as in other areas of natural language processing, though perhaps with less inclination for generic end-to-end architectures than in e.g. machine translation or named entity recognition.

In this section, we give an overview of the representations used for mentions (section 4.1.1), mention pairs (section 4.1.2) and mention clusters (section 4.1.3) used in coreference resolution systems.

4.1.1 Mentions

Given the formulation of coreference resolution as an operation on the set of mentions in a document (see section 2.2), a coreference resolution system needs as a bare minimum a way to represent mentions. In most existing systems, the representation of a mention is *local* —it only depends on its immediate context— and *autonomous* —it does not depend on the existence and characteristics of other mentions. However, some recent works such as K. Lee et al. (2018) and Kantor and Globerson (2019) include additional information (on preceding mentions and preceding clusters respectively) that depends on other mentions.

We make the assumption that a coreference resolution system has at least access to mention boundaries, but not necessarily *gold* mention boundaries. In practice, mention detection modules are usually not perfect and their outputs will contain some errors —both false positives and false negatives— and in an end-to-end system, they all have to be represented. Thus, for the remaining of this section, we will call 'mentions' all the elements that serve as inputs to a coreference resolution algorithm such as those described in section 4.2. Furthermore, some recent works on end-to-end systems, starting with K. Lee et al. (2017) do not even use a dedicated mention detection module and for them, 'mention' will encompass most of the spans in a document, though they usually use some filtering heuristics to reduce the combinatorial explosion.

External features

In most cases, when mentions are provided by a mention detection system, this system itself builds upon the outputs of a processing pipeline. In the prototypical case of BART (Broscheit

[1]Though not necessarily simultaneously: historically, this mostly concerned rules. Fully 'end-to-end' architectures that *do* only use characters stream-like inputs are a much more recent innovation.

et al. 2010; Versley et al. 2008), for instance, the main[2] mention detection pipeline includes a syntactic parser (from which POS tags can also be retrieved) and a named entities detector. In that case, this pipeline usually does not simply supply the mention boundaries, but returns mentions as objects with properties: intrinsic — such as a constituent type or a flexion — or structural — such as a position in a syntactic analysis or a semantic role.

We call these features *external features*, since they come from external resources that are usually not designed only for coreference resolution. We also include in these some basic features such as absolute positions in a document and character strings, that are used in similar ways but do not usually require extraneous processing as they are readily available in the raw source material.

In coreference resolution literature, most features are usually presented in terms of *agreements* and *distances* between their values for the mentions in a pair (see section 4.1.2), but defining them at the level of a single mention makes it clearer that they can also be used in other configurations, such as discourse-new detection or cluster-level representation (see section 4.1.3). Note however, that not all pair features can be meaningfully represented in this way, for instance the 'Appositive Feature' of Soon et al. (2001) models an actual linguistic relation and it makes more sense to consider it as a strictly pair-level feature.

Historically, the first widely used[3] set of features comes from Soon et al. (2001), which famously defined a set of 12 pair-level features that translate to the following mention-level features categories:

1. Position of the mention in the document, implemented as a sentence index.
2. Internal morphosyntax: constituent type (pronoun or noun phrase), presence of a definite article or a demonstrative pronoun, number and grammatical gender.
3. Semantic class among 'person', 'organization', 'location', 'date', 'time' 'money' 'percent' and 'object'. 'Persons' are also categorized in 'male' and 'female'. This is deduced from the WordNet (Miller 1995) entry for the head word of the mention.
4. The content of the mention itself as a string of characters. Articles and demonstrative pronouns are ignored.

This feature set had the advantages of being robust, relatively easy to compute (considering the then-current state of the art for preprocessing modules) and sufficiently small to allow extensive experimentations with machine learning architectures.

From then to 2006, subsequent works used more or less similar sets of features, with the notable exception of Ng and Cardie (2002) who extended it by introducing more semantic information extracted from WordNet as well as more ad-hoc heuristics such as a boolean feature for 'being part of a quoted string' (and more pair features, see section 4.1.2).

[2] An alternative pipeline replaces the parser with the combination of a chunker and a POS-tagger.

[3] Earlier rule-based systems and domain-specific machine-learning systems such as RESOLVE (McCarthy and Lehnert 1995) or Aone and William (1995) use similar mention representations, but in a less systematic way, are harder to generalize to other domains and were less directly influential in subsequent works, though it is clear that they did have an influence on Soon et al. (2001).

However, the addition of more features made the ensemble less easy to train, given the relatively small size of the learning material (the MUC corpus, MUC Consortium 1995c, 1998), with the need for non-negligible manual adjustments in order to actually improve over the Soon et al. (2001) baseline. This last issue was made clearer by the extensive studies of Uryupina (2006, 2007) on much larger sets of features: if linguistically-motivated handcrafted features coming from rich preprocessing pipelines do help, their efficiency seemed to be limited by the size (and sometimes consistency) of MUC. This assessment is also consistent with the findings of Culotta et al. (2007) and Bengtson and Roth (2008) whose systems obtained much better results with features sets quite similar to that of Ng and Cardie (2002) and without the extensive manual tuning needed for the latter, but using the much bigger ACE corpus (Doddington et al. 2004).

During this period, several works also showed an increased use of features associated with the *head word* of a mention. While Soon et al. (2001) and Ng and Cardie (2002) only used it as a way to probe WordNet, others such as Luo et al. (2004) and Uryupina (2003) started to use the POS of the head as part of mentions representation and Poesio et al. (2005) and Uryupina (2004) identified the crucial importance of head match as a pair feature for mentions-pair systems — reasserting a tendency known since Vieira and Poesio (2000) for noun phrase coreference detection.

This tendency is also confirmed by the quantitative study of Recasens and Hovy (2009), who found head match to be the most relevant of the common features used in coreference resolution even though it is irrelevant for pronominal reference. From a representation perspective, this means that the head word of a mention, along with its features, provides a good approximation of the mention as a whole — at least for most actual occurrences. Accordingly, all subsequent works using external features use head words as a central part of their mentions representations.

From then on, the representation of mentions used in coreference resolution systems mostly coalesced around the features used by Ng and Cardie (2002) (or subsets of them), including in languages other than English with Broscheit et al. (2010) and Kopeć and Ogrodniczuk (2012) using a compact set of such features for English, German, Italian and Czech. Attempts to expand this feature set to include deeper linguistic insights were also made. For instance, Haghighi and Klein (2009) propose to use heuristic estimations of salience and centering, inspired by their use in early works such as Hobbs (1986). These were also part of the highly successful dcoref system (H. Lee et al. (2013, 2011), also known as 'Stanford's sieve' and 'Stanford's deterministic coreference system'). However, Durrett and Klein (2013) shows that syntactic features such as salience can be approximated using simpler features (see section 4.1.1) provided that enough annotated data is available. Conversely, shallow semantic features such as those extracted from WordNet only provided marginal improvement when used for system mentions. As a result, their final system use a much shallower feature set, but relies instead on strong learning algorithms and a large dataset, a method that later works tend to follow as it reduces the need for costly expert rule handcrafting.[4]

In later years, with the rise of neural networks in natural language processing following Collobert

[4]It does enhances the need for annotated data and computational power, but at least for English, these are currently readily available.

et al. (2011), works using external features such as Clark and Manning (2016a,b) and Wiseman et al. (2016, 2015) mostly stuck with the feature sets from Durrett and Klein (2013), occasionally including some features proposed by Recasens et al. (2013). The —mostly successful— strategy behind this choice is to leverage the greater capacities of neural networks architectures to learn higher level representations from low-level features rather than relying on rich linguistic features.

Word-level features and compositions

The rich features described in section 4.1.1 have been shown to be efficient, especially in situations where annotated data is scarce, as in early efforts for English before the advent of large corpora or in current effort in lesser-resourced languages such as Basque (Soraluze et al. 2019). However, they all depend on the availability of a sufficiently reliable preprocessing pipeline for the target language *and domain*, whose quality strongly conditions that of the whole coreference resolution system.

When such a pipeline is not available, or when it is not accurate enough, the only features available are those that can be directly deduced from the contents of the mentions. As we mentioned in section 4.1.1, this content, represented as a string of characters, was already used in the representations of Soon et al. (2001) to fuel string-matching features (see section 4.1.2), and the content of the *head* word plays a central role in external features set (but depends on syntactic parsing). The first use of raw word-level features, however, came from Durrett and Klein (2013), who successfully introduced the contents of the first and last words of a mention and of the immediately surrounding words in their mentions representations.

When trained with enough data, these features can implicitly capture syntactic information in simple cases. For instance, in English, a subject NP is often followed by a verb and an object NP is often preceded by a verb and therefore in these cases, the words surrounding a mention can provided insights on its external syntactic features. Conversely, still in English, the first word of mention is often a good indicator of its definiteness —among other features— so a machine learning system can take advantage of the 'first word' feature to detect definiteness without help from a handwritten rule.

Still, no matter how useful these word-level features can be, Durrett and Klein (2013) and subsequent works using their feature set, such as Wiseman et al. (2015) or Clark and Manning (2015) were still relying on features targetting mentions' head words and were still dependent[5] on syntactic parsing for their mention representations needs. The reason was that while words at mentions' boundaries are informative, in general, they do not capture any semantic information about the mention and little information about its internal syntactic structure. Conversely, the head word of a mention often[6] carries the bulk of the semantic of the mention as a whole.

In that regard, the breakthrough is due to K. Lee et al. (2017): instead of using head words as

[5] A dependency that was not unjustified, since in an end-to-end setting, they also depended on syntactic parsing for mention detection in any case.

[6] But of course not always, consider for instance 'the chair of the department of social sciences' vs. 'the chair in the kitchen'.

rough summaries of mentions, they instead derive dense vector representations from the hidden states of a recurrent neural network operating on the words of the document. More specifically, given a mention, they apply a LSTM-based network (Hochreiter and Schmidhuber 1997) on the sequence of the words in its sentence, and use as a representation of the mention the hidden states of this network corresponding to the boundaries of this mention and a weighted average of those corresponding to the inner words of the mentions as a 'soft-head' (see section 5.2.2 for a more detailed description). These choices are reminiscent of the features used in previous works, but they have the crucial advantage of depending on all the words in the sentence and in the mention instead of a few heuristically chosen words. In fact, since this architecture is capable of processing whole documents[7] word-by-word, the only theoretical limits to the quality of its representations are the expressivity of its encoder (here the LSTM-based network) and the quality of its training. Most importantly, these representations do not require preprocessing beyond word segmentation, and are instead able to learn by themselves at least part of the features that classical pipelines provide: an examination of the internals of the models by K. Lee et al. (2017) suggest that the soft-head feature is often correlated with the actual syntactic head of a mention.

Additionally, with neural network models also came another powerful tool: word embeddings (Bengio et al. 2003), which provide dense vector representations for *words*. While word embeddings are almost always an improvement over sparse word representations for uses in neural network models (Collobert and Weston 2008; Collobert et al. 2011), *pretrained* word embeddings are particularly interesting for coreference resolution since they encode at least some lexical semantic information (Mikolov et al. 2013), giving to the systems that use them some help in fighting Durrett and Klein (2013)'s 'uphill battles'. Pretrained word embeddings are a form of external knowledge, but one that is more easily generalizable to other languages and domains that dedicated preprocessing pipelines, since they only require the availability of raw documents —for instance using large web crawl corpora (Ortiz Suárez et al. 2019)—, with no need for expert annotations. Their first use for coreference resolution is due to Clark and Manning (2016b) who used Mikolov et al. (2013)'s original word2vec pretrained embeddings, both for unique words (boundaries and head) and for sets of words using the averages of the embeddings of the context words of the mention (with left and right context windows of size 5), the content of the mention the whole document. Nitoń et al. (2018) uses similar word embeddings-based representations with success for coreference resolution in Polish.

Finally, in the last two years, the takeover of *contextual* word embeddings —based on techniques not too dissimilar to the LSTM-based feature extractor of K. Lee et al. (2017) and pretrained on language modelling tasks— have provided coreference resolution systems such as Joshi et al. (2019a,b), Kantor and Globerson (2019) and K. Lee et al. (2018) with word representations that directly include sentence-level information, obviating even the need for further recurrent word encodings.

[7] the limitation to sentence-level processing is a practical choice to limit the computational complexity of the model, not a theoretical one

Bound features

The features defined in section 4.1.1 are *autonomous*, in that they do not depend on the existence or the characteristics of other mentions in the document. Indeed, historically, representations involving several mentions have often been formulated at the level of pairs (see section 4.1.2) or sets of mentions (see section 4.1.3) and designed in reference to a particular coreference resolution algorithm. But the focus put by the use of neural networks architectures on rich representations rather than sophisticated algorithm have made *bound* representations —representations that do depend on the other mentions in a document— desirable, as their use can counterbalance the limitations of the greedy antecedent-finding algorithms popular in these systems.

Specifically, the 'higher order representations' introduced by K. Lee et al. (2018) and used by Joshi et al. (2019a,b) enhance mention embeddings by summing them with an average of the vector representations all preceding mentions, weighted by their coreference likelihood as estimated by a first round of their antecedent scoring algorithm, a process that can be iterated. Assuming that the coreference likelihoods are correctly estimated and that the initial representations are relevant, this average provides an approximate representation of the mention's antecedents.

Kantor and Globerson (2019) use a similar approach, but replace the mention pair formulation of K. Lee et al. (2018) by an entity-mention approach based on the fuzzy clustering proposed by P. Le and Titov (2017). In this formulation, fuzzy cluster representations are built recursively by estimating the degree of membership of every mention to the fuzzy clusters defined by the previous mentions initialized as singletons. In practice, this should be very close to what would be obtained with enough iterations of K. Lee et al. (2018)'s higher-order representations, but computationally more efficient.

4.1.2 Mention pairs

Mention pairs are critical objects in link-centric coreference resolution algorithms (section 4.2.2), but also in some entity-centric approaches such as Clark and Manning (2016b), where representations of *cluster* pairs is obtained by pooling the representations of the corresponding *mention* pairs. Historically, as we mentioned in section 4.1.1, mention pairs were actually the primary concept modelled for machine-learning coreference resolution systems, with mentions representations themselves being mostly latent. Conversely, in later works such as those following the approach of K. Lee et al. (2017), mention pairs are represented by a simple concatenation or pooling of the vector representations of the two mentions —leaving it to the antecedent scoring module to combine them in its internal layers to produce a score— but are not explicitly modelled, except for a single distance feature.

Except for systems that are more general anaphora resolvers, coreference resolution systems are only interested in mention pairs to determine if they are coreferent (or to compute a scalar coreference score), so the features used in their representations of mention pairs are *matching* features —boolean values indicating if the representations of the two mentions have the same value for a given feature— or *similarity* features —a degree of agreement between the two mentions on a given feature. These features are motivated by the (reasonable) assumption that

features of a mention are conditioned by properties of the corresponding entity, which implies that coreferent mentions should have matching features. This is also supported by empirical findings such as the preference for antecedent with identical syntactic functions (syntactic parallelism, Smyth 1994; Stevenson et al. 1995)

Therefore, a simplistic model could decide that a pair is coreferent if and only if all the features available for both mentions are matching. However, examples such as example 25 and corpus studies such as Antoine (2004) show that grammatical gender for instance can be mismatched between two coreferent mentions in at least some contexts, so that naïve 'all must match' heuristic is not sufficient in practice and combinations of features are needed.

(25) (a) *# At the farmhouse,* ***the cowgirl*** *left* ***his*** *lasso in the kitchen.*

(b) *At the Halloween party,* ***the cowgirl*** *left* ***his*** *lasso in the kitchen.* (Ackerman 2019)

Some systems also include *compatibility* and *preference* features, indicating if coreference between the two mentions is possible according to a given heuristic and its order of priority in case of ambiguity. These rules are usually derived from linguistic models of anaphora such as government and binding theory (Chomsky 1981) or empirical observations such as the preference for antecedent with subject roles. These features can also come from auxiliary tools, for instance Ng and Cardie (2002) use the output of 'a naïve pronoun resolution algorithm' as a pair feature.

Finally, most systems include *distance* features, integers or scalars encoding the distance between the two mentions. These might target various units, for instance Ng and Cardie (2002) include a distance in number of paragraphs, but the most common are distances in sentences, words and intervening mentions. These model to some extent the distance in discourse in terms of time or cognitive events, with the underlying supposition that discourse entities are stored in short-term memory, implying that coreference between distant mentions with no other coreferent intervening mentions follows different rules than coreference between close mentions. This property can be also be used to choose a single antecedent when several are equally likely, for example in the form of the 'closest-first' heuristic used by Soon et al. (2001).

4.1.3 Mention sets

As mention pairs for link-centric algorithms, mention *sets* lie at the core of most entity-centric algorithms. They model not only whole coreference chains, but also partial sets of coreferent mentions clustered together by a system as intermediary steps that can be seen as encoding the partial knowledge that a system can have of the corresponding entity at a certain step in the analysis of a document.

Accordingly, representations of mention sets that use explicit features tend to define them via unions, intersections or averages of mention-level features, meant to represent properties of the corresponding entities by inferring them from their known mentions. For instance, Luo et al. (2004) represent sets of mentions using multi-valued features that are the sets of the values of the corresponding mention-level features for the mentions that it contains allowing for lenient matching with other mentions: for an entity-mention pair (e, m) with $e = \{c_1, ..., c_n\}$, a given

feature is matching if there is at least one i such that the values of this feature are the same for m and c_i. In contrast, Yang et al. (2004) and later works such as Culotta et al. (2007) are more strict, and derive their set-level features by taking majority values over the mentions in a set, often with a majority ratio indicating if a feature value is shared by none, some, most or all of the mentions. These features provide sensible approximations of entity properties, but they can be too rough in practice, for instance for entities with varying properties or in end-to-end settings, where spurious mentions can blur their accuracy.

Björkelund and J. Kuhn (2014) propose a more precise representation by considering not features *sets* but features *sequences*, e.g. a set containing a proper noun, then a pronoun, then a common noun, then a pronoun would have a feature value for 'mention types' of 'NPCP'. By taking discourse order into account, these give a better accounting of the status of an entity, but the price of that precision is their far greater sparsity, making them hard to use in machine learning models.

Wiseman et al. (2016) tackle this sparsity issue elegantly, by using dense vector representations for clusters. Their system process documents left-to-right, in discourse order, and maintain a list of partial clusters. Every time a new mention is attached to an existing cluster, they derive a new representation for this cluster by applying an LSTM layer to the concatenation of its previous representation and the representation of the new mention (itself obtained from a neural network, see section 4.1.1). This approach relinquishes explicit control of clusters representations, but provides a relatively simple architecture for a system that process documents in their natural order and include effective clusters representations. A related[8] approach is used in the recent work of Fei Liu et al. (2019), where entities representations are stored in slots of a memory network (Sukhbaatar et al. 2015; Weston et al. 2015) and updated by a recurrent neural network at each affectation of a mention to an existing partial cluster.

4.2 Coreference resolution algorithms

In this section, we describe the existing algorithms for coreference resolution, the task of partitioning the set M of the mentions in a document in coreference chains. Many solutions exists for this task, differing by their model of coreference and their approach of M as either a sequence or an unordered set. Historically, the choice of a specific algorithm is mostly empirical, with no clear ranking between the solutions and no definitive study of their respective merits. Indeed, since comparisons are usually done on different representations, different supporting architectures and different material restrictions (for instance on computing power), which makes it hard to determine the properties of these algorithms from works on systems that use them.

4.2.1 Ordering

An important distinction between coreference resolution algorithms, is between *document-order* and *global* algorithms. *Document-order* algorithms build coreference chains by processing

[8] Although their model of coreference differs significantly

documents in linear reading order for written language or in utterance order[9] for spoken language, effectively modelling M as *sequence* of mentions. *Global* algorithms do not take this order into account and process documents globally, modelling M as an *unordered* set of mentions. Of course, systems using global algorithms are not necessarily (and indeed not usually) order-agnostic, but they encode order at the *representation* level (see section 4.1) rather than in their document processing.

This distinction is mostly independent of the model(s) of coreference used by algorithms. Some models are more naturally suited for one approach or the other —for instance, mention-centric models are usually used in document-order— but all models can in principle be used with or without order. Rather, the choice between these alternative is a trade-off between the ability to take all the information contained in a document into account and the available processing resources.

Generally speaking, document-order algorithms tend to be easier to develop (and to train, for machine-learning systems) and are less demanding in resources, since they can usually be formulated in terms of mostly *local* decisions and operate greedily. They are also more appealing from a modelling point of view: human processing of documents is by necessity done in order[10] and humans usually do succeed at coreference resolution, which implies that coreference chains in natural language documents must have structures that are compatible with and constrained by the structure of said documents. Therefore, a document-ordered algorithm, having direct access to this structure, could in principle use this structure to guide their operations. However, coreference resolution systems operate under more constraints than humans in terms of general language processing ability, world knowledge and explicit reasoning capabilities. Thus, a document-order algorithm might be unable to accurately process a document. In particular, compoundings of early errors make processing mentions near the end of a document particularly challenging.

Conversely, global algorithms can take into account more information, since they have access to the entirety of documents and can use this information to compensate for their lack of world knowledge. For instance, the grammatical gender of proper noun mentions is not necessarily easily accessible at the beginning of a document but might be deduced from later uses as in example 26, where without more context, 'she' might refer to 'Jean', but the later unambiguous reference of 'he' to 'Jean' makes far less likely.

(26) *Jean and Siobhán went to Orléans last year for an exchange program, and* ***she*** *was really happy about it. [...] Poor Jean,* ***he*** *was really unlucky.*

However, global processing of this kind is far more demanding in resources and more complex to design. As result, it is not uncommon to see global algorithms performing worse than document-order ones, which can make better use of the available resources by using better representations or a richer model of coreference.

[9] Which might be problematic for settings with significant overlapping, but no such resource for coreference is available yet.

[10] Although it can include some degree of delaying and backtracking. See for instance the results reported by Sturt and Lombardo (2005) for pronominal reference resolution.

4.2.2 Link-centric algorithms

Link-centric or *mention pair* algorithms operate within what we called the link-centric model of coreference in section 2.2. In this model, the objects of interest are links —potentially oriented— between mentions, and the coreference resolution task is cast as the task of finding a graph structure Γ on M such that the connected components of Γ are coreference chains. In this setting, the coreference chains are *latent* structures and are not explicitly modelled.

Formally, almost[11] all link-centric algorithms can be described by the two following elements[12]

- A pair scoring function $s : \mathrm{G} \longrightarrow \mathrm{V} \subset \mathbb{R}$, where $\mathrm{G} = \{\{m, m'\} \subset \mathrm{M} \,|\, m \neq m'\}$ is the set of unordered mention pairs
- A decoding procedure that maps the triplet $(\mathrm{M}, \mathrm{G}, s)$ to a graph (M, Γ) where $\Gamma \subset \mathrm{G}$, i.e. an edge selection heuristic

The decoding procedure is usually a heuristic optimization procedure operating on a constrained search space for Γ.

In that framework, coreference resolution is solved in two steps: first, scores are computed independently for all mention pairs, then a decoder selects the pairs in the output —either globally or in document order— to derive the system's output. The two steps are usually partially independent, in the sense that different decoders can be used with the same scoring function and vice-versa.

The simplest realization of this framework is used for instance by McCarthy and Lehnert (1995), where mention are classified are either coreferent or non-coreferent and coreference chains are build transitively. In this case, $\mathrm{V} = \{-1, 1\}$, $s(m, m') = 1$ if and only if m and m' are classified as coreferent (in McCarthy and Lehnert (1995), this was determined by a decision tree) and the decoding procedure globally selects all edges $\{m, m'\}$ such that $s(m, m') = 1$. In principle, this realization could yield perfectly accurate results, but it requires the classifier to have a perfect output, since every misclassification in the coreferent class will result in mistakenly merging two coreference chains. For MUC, which was the target of McCarthy and Lehnert (1995), this might yield good results, since the MUC metric is notably lenient with such overmerging (see section 2.4.2), but for most application this is an undesirable property, which is accordingly severely penalized by other metrics.

Antecedent-finding algorithms

Soon et al. (2001) also uses a decision tree classifier (albeit with different mention representations, see section 4.1.1), but proposes a different and important decoding procedure by selecting at most one antecedent per mention: a mention pair $\{m, m'\}$ such that m' is before m in document order is included in Γ if and only if m' is the closest[13] mention such that $s(m, m') = 1$, the so-

[11] To the best of our knowledge, only the twin-candidates model of Yang et al. (2003) requires a more complex formulation. Cai and Strube (2010) and Martschat et al. (2012) are formulated similarly but use higher level hypergraphs and multigraphs.

[12] This formulation is inspired by Lassalle (2015).

[13] This condition could also be included in s.

called 'closest-first' decoding. This decoding is much safer than the transitive closure decoding of McCarthy and Lehnert (1995): since every mention has at most one backward link, a spurious $\{m, m'\}$ link will only merge the chain of m' with the subsequent mentions of the chain of m instead of its whole chain. However, a spurious link for a discourse-new mention will still merge its whole chain.

This decoding procedure can be extended to the more general concept of *antecedent-finding algorithms*: algorithms where the search space for Γ is restricted to graphs where every mention is adjacent to at most one preceding mention in document order. It is easy to see that with this constraint, Γ is necessarily acyclic, but not necessarily connected[14] and therefore has a forest structure. Accordingly, we call these graphs *discourse-order forests*.

Of course, due to the structure of the search space, antecedent-finding algorithms are naturally discourse-ordered. While global alternatives such as an iterative search for a single coreferent mention m' for every mention in a document are theoretically conceivable, they don't really add much power if used greedily. Instead, the corresponding global approaches of latent trees usually uses more sophisticated decoding procedures (see 'Global decoding' below).

A limitation of the approach of Soon et al. (2001) is that it has to chose between two imperfect alternatives in a machine learning setting for the pairs $\{m, m'\}$ used as positive examples for $s(m, m') = 1$:

- Use only pairs where m' is the closest antecedent of m as positive examples, which restricts the size of the training set but is consistent with the decoding heuristic
- Use all coreferent $\{m, m'\}$ pairs, which gives more learning opportunity for the scoring function, but is inconsistent with the decoding, which only selects the nearest antecedent.

Soon et al. (2001) chose the first option with good results, but given the critical importance of a large training set for this kind of systems —highlighted by Durrett and Klein (2013)—, this choice is debatable.

Another obvious limitation of the closest-first heuristic is that it prevents easy resolution in some cases. Consider for instance the case where two identical proper noun mentions N and N′ appear in document order with an intervening pronoun P compatible with N′: N—P—N′. In that case, if the classifier deems that $\{\text{P}, \text{N}\}$ is coreferent, it will not even consider the $\{\text{N}, \text{N}'\}$ pair even if it far more likely to be coreferent.

A solution to these issues is proposed by Aone and William (1995) and Ng and Cardie (2002): rather than arbitrarily select the closest antecedent, the authors propose instead to select the antecedent for which their classifier has the most confidence, effectively switching from the previous binary-valued s to a scalar-valued one. By keeping the condition that $s(m, m')$ is set to -1 if $\{m, m'\}$ is classified as non-coreferent and to the confidence of the classifier else, this is equivalent to searching the discourse order forest Γ such that $\sum_{\{m,m'\}\in\Gamma} s(m, m')$ is maximal. This

[14] And indeed, a connected Γ is usually a bad result for a coreference resolution system. Note however that optimization algorithms operating on tree structures can still be used by modelling discourse-new mentions as linked to a dummy ε mention instead of having no antecedent

decoding procedure is known as 'best-first' clustering and tends to work better than closest-first clustering.

However, there is still an issue with the formulation of Ng and Cardie (2002) in that s is not directly optimized for in the training process, but is a by-product of training a binary classifier. More precisely, a binary classifier assumes independence between samples and its confidence score does not measure a degree of belonging to a given class. Rather, it is not unusual for binary classifiers to err with a high confidence score on unseen samples. Therefore, using a confidence score as an edge scoring function, if it does help in some cases, is not a good solution in general.

Bengtson and Roth (2008) and Denis and Baldridge (2008) offers an alternative architecture for best-first antecedent-finding, where s is directly learned.[15] This allows to take the relative scores of all the antecedent candidates of a mention into account, resulting in a more robust scoring function, while still keeping it strictly pairwise. This approach has been adopted by many subsequent works (Clark and Manning (2016a), Durrett et al. (2013), Joshi et al. (2019a,b), K. Lee et al. (2017) and Wiseman et al. (2015) among others), with various implementation and learning algorithms for s and various representations. Indeed, the case of K. Lee et al. (2017) in particular, which uses a rather straightforward learning algorithm and almost no additional knowledge suggests that despite its relative simplicity, this decoding procedure can be very efficient when used with solid input representations and a powerful learner for s. However, the architectures involved for these latter success are very demanding in data and in computing power and might not be applicable to situations with more constrained resources.

Global decoding

The performances of the antecedent-finding models are closely tied to the accuracy of the edge scoring function, and it is not always possible to improve it. Instead, another way to improve link-centric algorithms is to change the constrained document-order decoding for a global optimization procedure. By giving more power to the decoding procedure, these strategies can take advantage of all the information provided by the scoring function and compensate its eventual inaccuracies. As we mentioned in section 4.2.1, the price for these improvements is the need for more complex implementations and a higher computational cost.

One way to use a global decoding with a pairwise scorer is to formulate decoding as an Integer Linear Programming (ILP) problem, which returns a clustering with maximal total weight that satisfies arbitrary linear constraints. This is the choice made by Denis and Baldridge (2009), Klenner (2007), Roth and Yih (2007) and Uryupina (2010). The constraints can be formulated to enforce desirable properties on the resulting graph, such as *transitivity*, which ensures that the clusters generated by the decoding procedure do not include negative-scored links (i.e. non-coreferent mention pairs). The main issue with ILP formulation is that solving an ILP problem is NP-complete in general, and while Roth and Yih (2007) notes that ILP formulation of coreference are usually of a lower class of complexity, in practice, they are still too expensive

[15] Using a log-linear model for Denis and Baldridge (2008) and an averaged perceptron for Bengtson and Roth (2008).

for most uses (Lassalle 2015).

An alternative to ILP is the use of *latent tree* models (Björkelund and J. Kuhn 2014; E. Fernandes et al. 2012; Eraldo Rezende Fernandes et al. 2014; Lassalle and Denis 2015). Instead of considering all possible graphs in the search space for Γ, this formulation restricts it to the set of forest structures over M. This is a larger search space than in antecedent-finding algorithms — which considers only *discourse-order* forests — but it is still regular enough that polynomial optimization algorithms exist.[16] Another advantage of latent trees algorithms over ILP is that the scoring function can be optimized directly for the targetted structures, for instance using the structured perceptron algorithm (Eraldo R. Fernandes and Brefeld 2011), where the scoring function in ILP was trained for an antecedent-finding objective, which does not ensure that it would perform correctly with an ILP decoder.

4.2.3 Entity-mention models

As we have seen in section 4.2.2, despite their relative simplicity, mention-pairs models can provide good performances for coreference resolution. Yet, they require a choice between two extreme alternatives: document-order algorithms have limited access to the information contained in documents and require very accurate input representations and strong learners, while global algorithms have to deal with large and complex graphs of mentions which results in greater complexity. Between these alternatives, a middle ground is to use an explicit modelling of mention sets, which allows taking into account information beyond mentions without having to deal with overly complex structures on M. As we have seen in section 4.1.3, modelling mention sets is usually thought of as approximating entities, using the properties of the mentions in a partial coreference chain as an approximation of the attributes of the corresponding entity.

Mention-centric algorithms

The most common family of algorithms using entity modelling are mention-centric entity-mention models. These models correspond to the discourse-oriented modelling of reference that we have described in section 2.1.1: reference can be thought of as a (possibly directed) relation between a mention and an entity in a discourse model, either already existing or summoned by the mention. Accordingly, mention-centric models proceed to build coreference chains by iteratively considering every mention m and either attaching it to an already existing chain or creating a new chain (m). These algorithms are usually run in document order, which is also consistent with their closeness to human models of coreference resolution, i.e. mentions are considered in document order and are attached to sets of preceding mentions. In principle, a global mention-centric model would be conceivable although harder to relate to theoretical principles, but no work has currently explored this possibility.

Conceptually, this is very close to antecedent-finding, except that the antecedent here is not a single mention, but a set of mentions carrying higher level information. Theoretically, this could have the advantage of reducing the number of alternative antecedents, however, many

[16] Such as Edmonds' algorithm (Edmonds 1967) for directed forests or Kruskal's algorithm (Kruskal 1956) for undirected forests.

implementations still consider attachment to a cluster as an attachment to a specific mention in a cluster and thus describe the whole procedure as antecedent mention finding. This distinction does not make a fundamental difference, since it can be seen simply as a different procedure to choose an antecedent cluster.

A clear distinction with antecedent-finding models is that while choosing an antecedent for a mention is independent of the choices for the other mentions, since the available mention sets available for attachment at any single time depend on all the previous decisions. This means that the result of a mention-centric algorithm depends on a chain of dependent attachment actions rather than on independent antecedent-finding, making learning the full problem intractable. In practice, as with other such problems in natural language processing, this is usually solved by using a Markovian model where the choice of action at each time step depends only on the current state (in that case the list of incomplete clusters) rather than the full sequence of actions. During learning, the attachment choice is usually presented with gold clusters from an oracle.[17] At inference time, decoding is usually done either greedily or using a beam search.

Historically, the first mention-centric models came simultaneously from Luo et al. (2004) and Yang et al. (2004), with a similar formulation as an antecedent-finding procedure with cluster-level features using classifiers, Luo et al. (2004) with an interesting formulation of beam search using the Bell tree structure to model incremental partition building and Yang et al. (2004) with a more usual greedy decoding. Recasens and Hovy (2010) also uses a similar model (CISTELL) to investigate discrepancies in coreference resolution across corpora and metric bias. Rahman and Ng (2011) proposes a rather natural extension of the previous entity-mention models to act by *ranking* clusters, similarly to the move from mention-pair classification to antecedent scoring models.

Most recently, Wiseman et al. (2016) shows that using neural network architectures and dense representations could greatly improve the performances of mention-centric models, consistently with the improvements they provide to antecedent scoring models (Wiseman et al. 2015). The memory network architecture of Fei Liu et al. (2019) is also a form of entity-mention network which shows encouraging performances for pronominal reference resolution.

Entity-centric algorithms

A dual approach to the mention-centric algorithms where coreference chains are built by attaching mentions and the base objects are *mentions* is to consider *coreference chains* as the base object and to build them by agglutinating partial clusters in multiple passes. This has the crucial advantage over other iterative methods that failing to detect that two mentions are coreferent at first can be recovered from at any time by merging their clusters, possibly using cluster-level information. In contrast, in antecedent-finding and mention-centric model, setting a mention as discourse-new is a final decision that can not be overridden by later information.

This approach was originally proposed by Raghunathan et al. (2010), using a rule-based system in the so-called 'multi-sieve' approach also used in the later works of H. Lee et al. (2013, 2011,

[17] To our knowledge, while the bias induced by static oracles are well known for other tasks, dynamical oracles, used for example in syntactic parsing (Goldberg and Nivre 2012), have never been used for coreference.

2017). In this formulation, a systems starts with a partition of M in singletons and applies successively a series of rules (handcrafted in H. Lee et al. (2013, 2011) and partially learned in H. Lee et al. (2017)) to merge existing clusters in document order.

A similar cluster agglomeration model is proposed by Stoyanov and Eisner (2012), but uses a reinforcement learning approach operating in global fashion: at every step of the algorithm, an agent observes the current clustering and emits an action to either merge two clusters or stop the process. This formulation allows performing 'easy' cluster merging first and delaying complex decisions until later, at which point the information contained in the partial clustering will hopefully make them easier. This approach is also used by Clark and Manning (2015, 2016b), which extract cluster-pair representations from auxiliary link-centric models and in the case of Clark and Manning (2016b), dense mention representations and a deep policy network.

4.3 Mention detection

Mention detection is a crucial part of an end-to-end coreference resolution system and the results on coreference resolution presented in the recent literature tend to report end-to-end results —following the standard tracks of the SemEval and CoNLL shared tasks (Pradhan et al. 2012, 2011; Recasens et al. 2010). In this end-to-end setting, the accuracy of mention detection is also crucial for the overall results, for instance, Ng (2008) reports far worse results for their unsupervised system when using system instead of gold mentions, an observation that is shared in the systematic study of Stoyanov et al. (2009).

Despite this, the attention received by the mention detection task has been mostly anecdotal: few works are specifically dedicated to mention detection, and it is rarely studied in details in works on end-to-end systems. This lack of interest might come in part from the particular status of mention detection in resources for English: among the corpora used in shared tasks, viz. MUC (Hirschman and Chinchor 1998; MUC Consortium 1995b), ACE (Doddington et al. 2004), Ontonotes (Pradhan et al. 2007a), SemEval (Recasens et al. 2010) and ARRAU (Poesio and Artstein 2008; Uryupina et al. 2019), only the last two includes annotations for all mentions including singletons. Furthermore, the English part of SemEval has not received as much attention as it could have, since it was mostly supplanted for English by the Ontonotes-based CoNLL corpora (see section 3.1.4), and the use of ARRAU in a shared task is very recent (Poesio et al. 2018).

Works with specific studies of mention detection have mostly been focused on ACE, which does include singletons, but only for the restricted class of entities that are annotated, and these works also include mention type classification, which makes it hard to assess their general results on untyped detection. For corpora that do not include singleton, mention detection is hard to study, since it is compounded with coreference resolution. Consequently, in corresponding works, mention detection is often a mere pre-filtering recall-oriented step and the techniques used to perform it are not particularly concerned with the precision of the detection, making simple heuristic mostly sufficient. In fact, the success of the span-based models initiated by K. Lee et al. (2017) suggests that for these resources, an explicit mention detection step is not an obligation.

It is hard to determine precisely the origin of mention detection in coreference resolution, with early works on MUC using gold mentions or simply all NPs extracted from parse trees. Most of the works that give any details on their procedure seem to use rule-based techniques such as those reported by Uryupina (2010) or H. Lee et al. (2011): NPs, pronouns and named entities are extracted, either using gold knowledge or from preprocessing pipelines, and a set of filters is applied on them to filter out those that are known to be non-referential (which is usually corpus-dependent, as noted by Stoyanov et al. (2010), whose RECONCILE system use different sets of rules depending on the resource it is used on). When syntactic parsing is not available, some works such as Soraluze et al. (2012, 2017) also report encouraging results by combining lower-level preprocessing tools with extensive rule-based systems.

Beyond rule-based systems, Uryupina and Moschitti (2013) propose a machine learning system, still using syntactic parsing, but learning a noun-phrase classifier to filter out non-referential NPs. This allows extending to new languages more easily than when using language-specific rules, with a significant improvement over the all-NP baseline previously used by many systems for the Arabic and Chinese portions of Ontonotes.

More recently, the very promising work of B. Wang et al. (2018) describes a method for a full end-to-end mention detection system in the nested named-entity recognition task that does not rely on pre-existing parse trees, but learns instead a partial transition-based parser in a semi-supervised setting where the only supervision is on the correct detection of mention spans.

4.4 Systems for French

Due to the recency of publicly available annotated resources for coreference resolution in French (see section 3.2), very few works exist on coreference resolution for French. Two rule-based systems, Longo (2013) and Trouilleux (2001), for narrow coreference resolution exist, but they suffer from a lack of resources, both for development and evaluation.

Such resources were first provided with the release of ANCOR (Muzerelle et al. 2013a, 2014), but still, prior to this thesis, only two works studied coreference resolution, perhaps due to the very specific nature of this corpus and the lack of suitable preprocessing tools.

Désoyer et al. (2015a) proposes the first coreference resolution system for French: CROC. It was not meant as a full end-to-end system, since it only performed coreference resolution using both gold mention spans and the gold mention features described in section 3.2.2. The coreference resolution model use *mutatis mutandis* the same representations as Ng and Cardie (2002) and the same closest-first antecedent-finding algorithm as Soon et al. (2001), supported by a SVM classifier for their best model. Their results (reported in table 5.5, p 90) were very encouraging, considering the relative simplicity of the system, which confirms the relevance of the gold mention features included in ANCOR for coreference resolution.

The more recent work of Godbert and Favre (2017) is an attempt at an end-to-end coreference resolution system, but was in this work only limited to pronominal reference and so-called 'direct anaphora' between NPs. In contrast with Désoyer et al. (2015a), it does not use a machine-learning system, but a rule-based system on top of the MACAON pipeline (Nasr et al. 2011),

with a strong focus on pronouns and a rather simple heuristic for noun phrases.

Chapter 5

Automatic coreference resolution

Despite the large body of work in natural language processing for French in general, the specific task of coreference resolution has not received as much attention as it has in other well-resourced languages (such as Czech, Polish, German, Spanish or Catalan). Indeed, without any large-scale publicly available corpus with coreference annotations, until the publication of ANCOR (Muzerelle et al. 2014), early attempts (Longo 2013; Trouilleux 2001) at building a coreference resolution system were largely limited by either the scarcity of resources or their usage restrictions. Since then, some partial proof-of-concept systems (Désoyer et al. 2015a; Godbert and Favre 2017) have been produced and some experiments (Brassier et al. 2018) have been conducted using ANCOR, but there is yet no publicly available end-to-end coreference system for French.

In this chapter, we propose to address this lack, by proposing DeCOFre: an end-to-end coreference resolution system leveraging state-of-the-art neural architectures and trained and evaluated on ANCOR.

In section 5.1, we list our requirements for an end-to-end coreference resolution system and justify our choices to designing a new system (instead of adapting an off-the-shelf system to French) and to use a two-step pipeline architecture as opposed to the fully end-to-end architecture used in the current state-of-the-art.

In section 5.2, we expose the structure of our system from a global point of view, describe our implementation in details and motivate our architectural choices.

In section 5.3, we describe the optimization process used to train the parameters of our system on ANCOR.

Finally, in section 5.4, we detail the experimental settings we used to assess the performances of our system and report some additional experimental results, meant as partial empirical validations for some technical choices described in the previous sections.

5.1 System scope

In designing our coreference resolution system, given our goal of end-to-end coreference resolution on ANCOR, in the configuration described in section 3.2.2 our main requirements for it were that

1. It had to be able to work directly on raw transcriptions and detect mentions by itself.
2. It had to treat singleton as chains.
3. It had to be *suitable for spontaneous spoken language*, which can be significantly different both from written language and from controlled and planned spoken language. The most inconvenient of these differences for us were the lack of clearly defined sentences boundaries and the presence of distractors such as disfluencies, incomplete words and phatic markers.

We also had some secondary requirements

4. The system had to be suitable for our data —namely ANCOR— with its specificities beyond language and genre. Most notably its great heterogeneities in terms of document sizes and natures (see section 3.2.2) was an issue.
5. The lack of pre-existing NLP resources for spoken French: at the time of writing, there are no publicly available systems for the usual preprocessing steps used in coreference resolution (POS-tagging, parsing, NER…) that reliably supports spontaneous spoken French. Moreover, systems designed for written French are of unreliable quality on spoken French and would require considerable post-correction work to reach the quality required for direct use (see Grobol et al. (2017b) and Tellier et al. (2014), ways of partially overcoming this issue are explored in chapter 6).
6. The limitations of our hardware in terms of memory and processing power. For instance, our early attempts to use K. Lee et al. (2017)'s E2E system directly on ANCOR simply crashed, having filled up all the memory available on our machines.

Considering these points, the publication of the E2E[1] system K. Lee et al. (2017) was determinant, as it provided us with an architecture that met our requirements 1 and 5 by being able to take raw text as input, without relying on external resources while still reaching sate-of-the-art quantitative results on the CoNLL-2012 benchmark.

However, while making E2E comply to 2 could have been done, for instance by using their s_m score or a learned $s(\cdot, \epsilon)$ score, as a discourse-new detector, the interaction between 6 and 4 prevented us to use it as a base system to build upon.

Indeed, since E2E always operate at the document level, whole documents have to be kept in memory during training and error backpropagation must take into account the operations performed on all of their text spans. To counter the prohibitively high computational and

[1]For 'end-to-end neural coreference resolution system'. The authors did not give a name tho their system, but this abbreviation is becoming common in recent works.

memory cost of this approach, K. Lee et al. (2017) resort to a variety of aggressive pruning at every step such as:

1. Truncating documents (to a maximum of 50 sentences)
2. Restricting the maximum span width and number of spans per word
3. Restricting the number of antecedent candidates

2 is particularly problematic as it is done *during training*, with the discarded spans contributing nothing to learning at the current step, and thus slowing down the training process. Furthermore, it implies that the training process must use whole documents as batches, preventing the use of common training techniques such as mini-batching or sample shuffling, which results in a greater reliance on the fine-tuning of the training hyperparameters.

At the root of these issues is the lack of an explicit mention detection step in E2E, which would single-handedly:

- Reduce the cost of antecedent-finding
  - At inference time, by considerably reducing the number of alternatives to consider
  - At training time, by removing the reliance on heavy pruning and truncations
- Makes the system more modular, which would help for both downstream applications and evaluation
- Allows working with singleton entities

while making the whole architecture easier to implement.

Therefore, in order to develop a system for end-to-end coreference resolution for oral French, our choice is to design a classical two-step pipeline for coreference resolution using and adapting the neural building blocks and the span-based paradigm of K. Lee et al. (2017).

5.2 The DeCOFre system

5.2.1 General architecture

In its end-to-end inference mode, given a raw text seen as a sequence of tokens $T = (w_1, ..., w_n)$, DeCOFre works according to fig. 5.1, which is essentially a combination of mention detection by span classification and antecedent scoring with greedy document-order decoding (see section 4.2).

At the end of this process, the (M, A) graph has a forest structure, whose connected components (viz. trees) are the inferred coreference chains.

5.2.2 Representing text spans

The core of DeCOFre is the g span representation function, implemented as a neural network N_g. It is similar in its main architecture to the **g** representation from K. Lee et al. (2017), made

Figure 5.1: DeCOFre inference mode operating process

```
1  Let M = ∅                                   the set of the currently detected mentions
2  Let A = ∅                                   the set of the currently detected coreference
                                               links
3  For i from 1 to n
4      For j from i to n
5          Let m = (w_i, ..., w_j)             the (i, j) span
6          Compute g_m = g(m, ψ(m))            its representation
7          Determine d(g_m)                    its class
8          If d(g_m) is a mention class
9              Compute s_n(g_m)                the confidence that m is discourse-new
10             For c in M
11                 Let a(m, c) = s_a(g_m, g_c, ϕ(m, c))   the confidence that c is an antecedent of m
12             Let c = arg max_{x∈M} a(m, x)   the most likely antecedent
13             If a(m, c) > s_n(g_m)
14                 Add (m, c) to A
15             Add m to M
```

ψ and ϕ are additional features functions.

more general by removing the dependency on *a priori* sentence segmentations of documents and making the influence of the immediate context of text spans more direct. It also includes some improvements inspired by other recent works on neural architectures for natural language processing tasks.

Input representations

As mentioned earlier, we consider documents as sequences of words (or tokens). In order to provide our neural architecture with efficiently usable representation of these, we use static *word embeddings* (Bengio et al. 2003). For this, we build a fixed dictionary mapping the words occurring in the considered training set, excluding hapax, which make up about 40 % of the total vocabulary —consistently with Zipf's law (Zipf 1949). This has the double advantage of avoiding the addition of hard to train representations to the model and making it more robust to out-of-vocabulary words.

Formally, we note $e_w(w_i)$ the word vector corresponding to w_i..

However, there is information in inputs that is not easily accessible from word embeddings. Even in the case of perfectly standard language, typos and out-of-vocabulary words are all lost if we only rely on word embeddings. This issue is particularly salient for transcriptions of spontaneous speech, where incomplete, garbled and rare words are frequent, as in example 27.

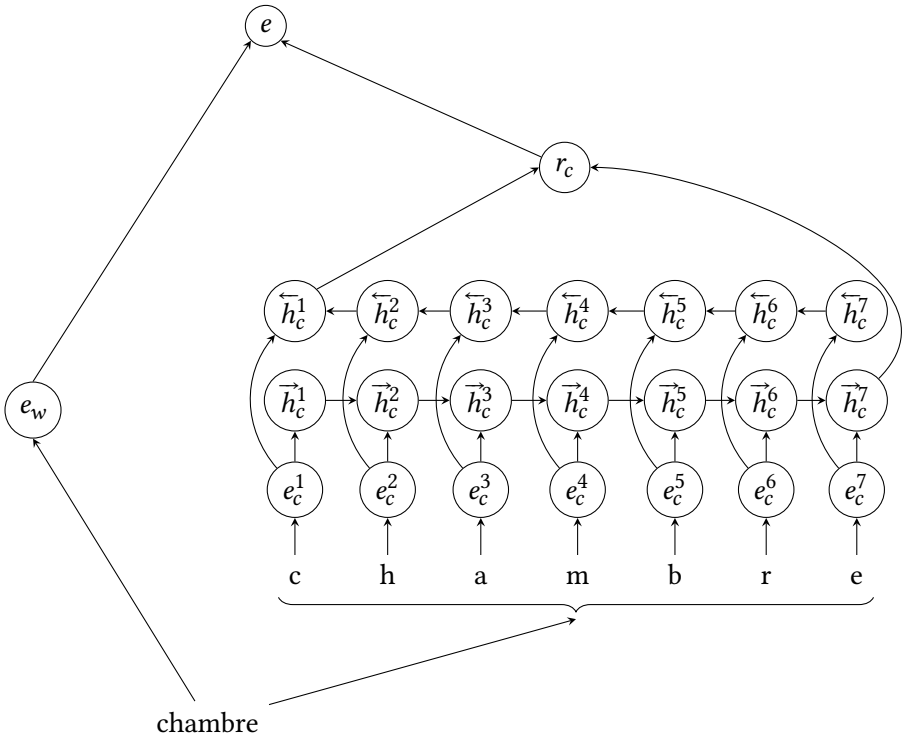


Figure 5.2: Intrinsic words representations in DeCOFre

(27) *bah j'aime euh j'aime bien les disons les euh le ciné- le cinéma actue l euh le cinéma euh euh un petit symbolique*

(well I like uh I rather like the say the uh fil- contemporary fil- uh films that are uh uh a small metaphorical) (Muzerelle et al. 2014)

To address this, following Ling et al. (2015) we also use character-level representations of words *via* the r_c, function implemented as a recurrent neural network —in our case, as in **dinarelli2019HybridNeuralNetworks**; Dinarelli and Grobol (2018), a bidirectional GRU layer (Cho et al. 2014), which is theoretically less powerful than the LSTM layers used in some other works (Weiss et al. 2018) but seems sufficiently efficient in this case as poor man's morphological analyser.

Formally, a word w_i made up of the characters $c_1, ..., c_n$, is represented at the character level by a vector $r_c(w)$

$$\begin{aligned} \mathbf{e_c} &= (e_c(c_1), ..., e_c(c_n)) \\ \overrightarrow{\mathbf{h_c}} &= \overrightarrow{\mathrm{GRU}_c}(\mathbf{e_c}) \\ \overleftarrow{\mathbf{h_c}} &= \overleftarrow{\mathrm{GRU}_c}(\mathbf{e_c}) \\ r_c(w) &= \mathrm{FFNN}_c\left([\overrightarrow{h_c^n}, \overleftarrow{h_c^1}]\right) \end{aligned} \quad [5.1]$$

Where $e_c(c_i)$ is a character embedding corresponding to c_i, FFNN_c is a feedforward layer and $\overrightarrow{\mathrm{GRU}_c}$ (resp. $\overleftarrow{\mathrm{GRU}_c}$) is the forward (resp. backward) direction of the bidirectional GRU layer.

As in Ballesteros et al. (2015) and Plank et al. (2016), where the best results for dependency parsing and POS-tagging respectively were obtained by using both word- and character-level information, we derive from this our final intrinsic word representations by concatenating word embeddings and character-level representations

$$e(w_i) = [e_w(w_i), r_c(w_i)] \quad [5.2]$$

Contextual words representations

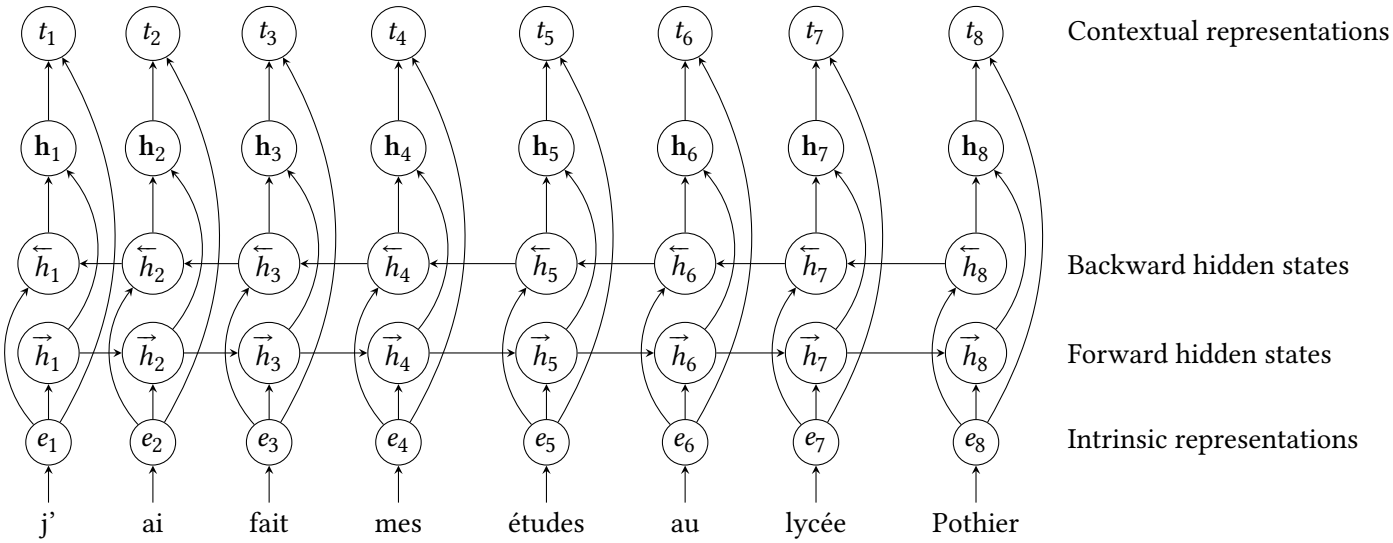


Figure 5.3: Contextual words representations in DeCOFre

Since our objective is to build a span representation module, we could stop at intrinsic words representations. After all, as noted by Shen et al. (2018), simple pooling techniques work at least as well as more complex recurrent or convolutional architectures for a variety of semantic task across several datasets.[2] For coreference specifically, Nitoń et al. (2018) shows that state-of-the-art results can be obtained for Polish in the gold mention setting with a simple average pooling mechanism on word embeddings (but using additional knowledge in the form of pair features).

However, there are reasons to think that what essentially amounts to a bag-of-words approach might be too limiting of the expressiveness of a span representation module. Indeed, research on distributed words and segment representation in the last years tend to show that representations able to take word order and word contexts into account are vastly more performant than the alternative for most NLP tasks (Devlin et al. 2018; Peters et al. 2018) including coreference resolution. (K. Lee et al. (2018), see also section 6.2.2). This is not to say that these findings are a definitive optimal solution to the problem of representing human languages in neural networks: even in more sophisticated —and indeed state-of-the-art for most natural language understanding tasks[3] — architectures that take word order into account, recent results (McCoy et al. 2019; Niven and Kao 2019) suggest that neural network systems might fall into simple heuristics that are too limited for reliable coreference resolution. That said, it seems safe to

[2]Even if these findings were limited to English and their applicability to other languages is not clear.

[3]Though here, too, the reasons behind these results are still not completely clear.

assume that these representations should be more suitable for our tasks than order-agnostic representations based purely on pooling word-level representations.

Considering this, we chose to imbue our word representations with contextual information, by augmenting them with the hidden states of a recurrent neural layer applied to each span, while taking into account its left and right contexts. In our baseline, this layer is a Bi-LSTM (Graves and Schmidhuber 2005; Schuster and Paliwal 1997), chosen for its superior computational capabilities (Weiss et al. 2018) over GRU.

Another possible choice would have been an attention-based architecture, with the most serious candidate being the Transformer (Vaswani et al. 2017), but we decided against it. This choice is motivated by their novelty and lack of applications beyond machine translation at the beginning of this work, while the efficiency of the LSTM units was supported by more empirical knowledge.[4] Since then, other works suggest that attention-based sentence modelling architectures for span representation could be performant for coreference resolution (Joshi et al. 2019b; Webster et al. 2018) and one work in particular (Joshi et al. 2019a) that they could be far more efficient than previous approach, although in this case, as with most of the recent achievements in the wake of (Devlin et al. 2018), the relative contributions of architectures, data, training and model size are yet to be properly disentangled (Rogers 2019).

Specifically, given a text T $= (w_1, ..., w_n)$, to compute the representation of span $(w_i, ..., w_j)$, we truncate T to a fixed context window clipped at utterance boundariesS, formally

$$\mathrm{S} = (w_{\max(i-ctx_r,1)}, ..., w_i, ..., w_j, ..., w_{\min(j+ctx_r,n)}) \tag{5.3}$$

where the positive integers ctx_l and ctx_r are hyperparameters of the model.

We then apply a Bi-LSTM layer on the context-free words representations $\mathbf{e} = (e(w))_{w\in \mathrm{S}}$ and keep its hidden states $\mathbf{h} = (h_k)_{k\in \mathrm{S}}$ computed as

$$\begin{aligned} \overrightarrow{\mathbf{h}} &= \overrightarrow{\mathrm{LSTM}}_w(\mathbf{e}) \\ \overleftarrow{\mathbf{h}} &= \overleftarrow{\mathrm{LSTM}}_w(\mathbf{e}) \\ \mathbf{h} &= \left([\overrightarrow{h_k}, \overleftarrow{h_k}]\right)_k \end{aligned} \tag{5.4}$$

where $\overrightarrow{\mathrm{LSTM}}_w$ (resp. $\overleftarrow{\mathrm{LSTM}}_w$) is the forward (resp. backward) direction of the Bi-LSTM layer.

Finally, to coax our model into learning better context-free representations instead of relying entirely on the recurrent layer, we add a skip connection around it by concatenating $\mathbf{h}$ with $\mathbf{e}$ Our final contextual representation for a word w_k is then

$$t_k = [h_k, e(w_k)] \tag{5.5}$$

Span embeddings

Span embeddings are derived from the contextual words representations in three different ways

[4] An inkling connection between these two families of architectures can be found in Levy et al. (2018), which could explain why both work so well on similar tasks

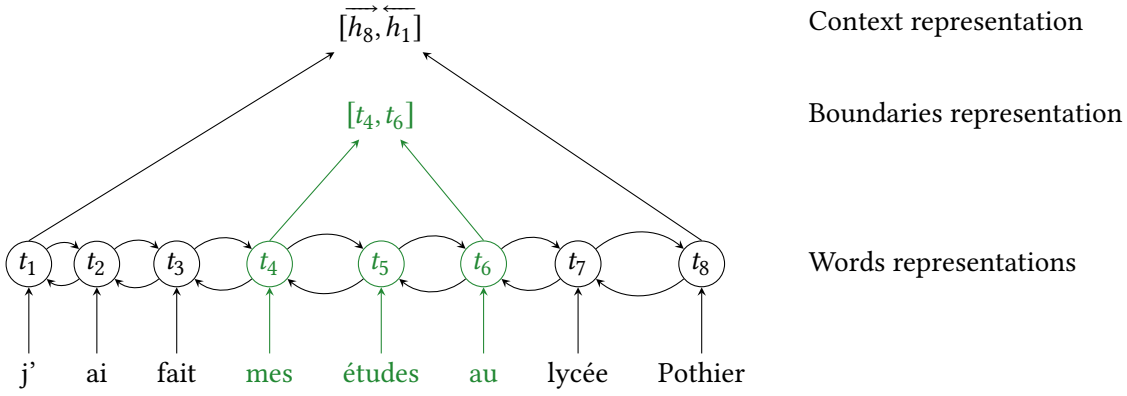


Figure 5.4: Boundaries and context embeddings in DeCOFre

Boundaries The simplest part of the span embeddings consists of the representations of the boundaries of the span, that is, the representations of its first and last words. This can be seen as a form of (gated) recurrent segment representation (Cho et al. 2014; Sutskever et al. 2014; Vinyals et al. 2015) generalized to the bidirectional case by considering the hidden states of the forward and directions for both the first and the last word of the segment instead of using only a single boundary and a single direction. This is mainly motivated by our objective of getting the mentions with their exact boundaries rather than a more relaxed target.

Consistently with K. Lee et al. (2017) we derive this representation from the concatenation of the representations of the boundary tokens. Formally, this is $b_{i,j} = [t_i, t_j]$.

We decided against the earlier LSTM-minus approach introduced by W. Wang and Chang (2016) and used by Cross and Huang (2016) and Stern et al. (2017), which yields representations of lower dimensions but would impose a much more constrained structure to the LSTM hidden states *and* the intrinsic word representations in our architecture.

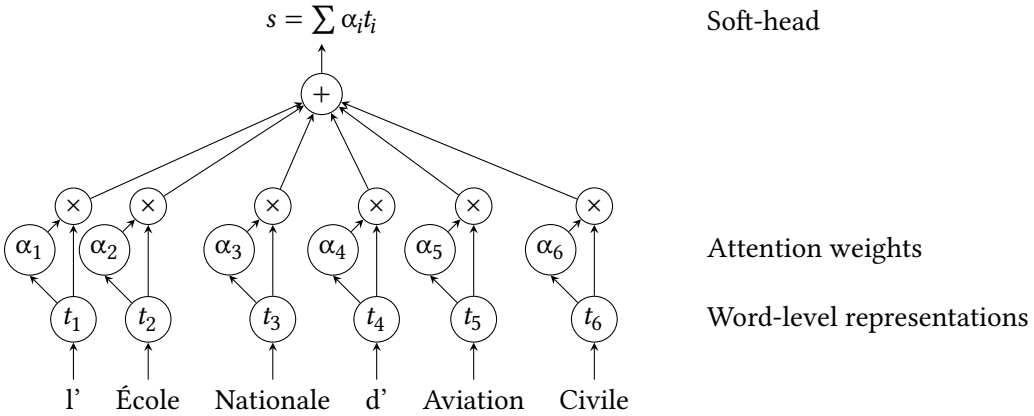


Figure 5.5: Soft-head self-attention in DeCOFre

Soft-head The second part of the span embeddings is a structured self-attentive embedding (Lin et al. 2017) of the span. Its use on spans rather than full sentences (as in the original from Lin et al. (2017)) was motivated by K. Lee et al. (2017) for its similarity to the syntactic head features used by previous systems, to which it provides a substitute in a context where syntactic parsing is either unavailable or unreliable.

Formally, we represent the span $(w_k)_{i\leqslant k\leqslant j}$ by a weighted average $s_{i,j}$ of the contextual representations $(t_k)_{i\leqslant k\leqslant j}$, where the weights $(\alpha_k)_{i\leqslant k\leqslant j}$ are themselves computed from t_k, that is

$$\begin{aligned}(\alpha_k) &= \underset{k}{\text{softmax}}(\text{FFNN}_\alpha(t_k))\\ s_{i,j} &= \sum_k \alpha_k t_k\end{aligned} \tag{5.6}$$

According to K. Lee et al. (2017) the maximal α_k is usually given to the syntactic head of a span or to its most semantically significant word in cases where the head does not bear a lot of meaning (as is the case for coordinations for example). In our case, since the syntax of spoken French tends to be more irregular,[5] it is interesting to be able to attend to several words. In principle, this could be learned by FFNN_α in this formulation, but this is made unlikely by the high concentrating effect of the softmax normalization, which by design tends to concentrate the overwhelming majority of the weights on the maximal coordinate. To allow more expressivity, we follow the example of Lin et al. (2017) and concatenate *several* such self-attentive embeddings to obtain s. Formally,

$$\begin{aligned}(\alpha_k^1) &= \underset{k}{\text{softmax}}(\text{FFNN}_\alpha^1(t_k))\\ &\vdots\\ (\alpha_k^\tau) &= \underset{k}{\text{softmax}}(\text{FFNN}_\alpha^\tau(t_k))\\ s_{i,j} &= \left[\sum_k \alpha_k^1 t_k, \ldots, \sum_k \alpha_k^\tau t_k\right]\end{aligned} \tag{5.7}$$

where the *number of heads* τ is an hyperparameter of the model.

Context Finally, we also use representations of the contexts of the spans, in the form of the final states of $\overrightarrow{\text{LSTM}}_w$ and $\overleftarrow{\text{LSTM}}_w$, that is $ct_{i,k} = [\overrightarrow{h_{n+p}}, \overleftarrow{h_{-\ell}}]$. Indeed, while information on the span's context outside $(w_k)_{i\leqslant k\leqslant j}$ might flow to $b_{i\leqslant k\leqslant j}$ and $s_{i\leqslant k\leqslant j}$ via the recurrent parts of $(t_k)_{i\leqslant k\leqslant j}$, it is more straightforward to provide it directly to the subsequent parts of our architecture, especially since in our implementation it comes with no additional computational cost.

The span embedding $\hat{g}_{i,j}$ is computed from the concatenation of all of these parts by a feedforward layer, that is

$$\hat{g}_{i,j} = \text{FFNN}_o([b_{i,j}, s_{i,j}, ct_{i,j}]) \tag{5.8}$$

Final representation To $\hat{g}$, which is entirely derived from the surface form of the spans without any supervision other than the architecture, we add some explicit features by concatenating some non-pretrained low dimensional feature embeddings to a vector ψ. For the

[5] Syntactic constructions are usually shallower than in written languages, but can be much wider and tends to use less explicit compositions, such as parataxis. (Lacheret et al. 2014)

feature-poor model presented in this chapter, ψ simply consists of the length of the span (see section 6.1).

The final representation of a span g_m is then the concatenation of $\hat{g}(m)$ and the feature embedding $\psi(m)$, projected via a linear function π_s to keep a constant dimension for the subsequent layers independently of the number of features used and of their embedding dimensions.

$$g_m = \pi_s([\hat{g}(m), \psi(m)]) \quad [5.9]$$

5.2.3 Mentions detection

Among the different mention detection strategies described in section 4.3, the resource-poor setting of DeCOFre ruled out the syntax-based methods (due to the lack of reliability of the available parsers), and made the label+transform method proposed by Soraluze et al. (2012, 2017) very inconvenient. Therefore, we were left with a choice of a transition-based method (à la B. Wang et al. (2018)), which while very promising requires a complex machinery to train, or our final choice: a span based classifier.

This last method has the advantage of being very easy to implement, as it simply uses a neural classifier —a relatively well-known model— on top of the span embeddings described in section 5.2.2 that we intended to use for the coreference resolution subtask in any case. Therefore, in DeCOFre, mention detection is formulated as a classification task on the set of all text spans.

The theoretical complexity of this formulation for a document of n words is $O(n^2)$, but in our case, since mentions can't cross utterance boundaries, it is reduced *in average* to $O(np)$ where p is the average sentence length. This can be sufficiently small in some contexts, where utterances lengths are limited.

However, since the utterances in ANCOR can have a length of several hundreds of words (up to 542) and mentions are most a few words long (less than 5 % consist of more than 4 words and the longest mention is a mere 17 words long), we add a further restriction of considering only spans of length inferior to an integer $\ell_{\max}$, bringing the number of spans to consider in the worst case to $O(n\ell_{\max})$.

In our implementation, the classification function d is a feedforward neural network that returns non-normalized class scores. In principle, this could be done by a simple binary classifier, but since ANCOR provides mention types information (see section 3.2.2), we chose instead to classify the spans between pronouns, noun phrases and non-mentions.

Formally, this is then

$$d(g_m) = \arg\max(\text{FFNN}_d(g_m)) \quad [5.10]$$

5.2.4 Coreference resolution

As seen in section 4.2, there are several competing algoritms to partition a set of mentions in coreference chains. Among these, link-centric algorithms have the advantage of being the most lightweight, since they only require representing two type of objects: mentions and mention

pairs themselves, without having to deal with the extra level of entities. In our case it is even most interesting, since it allows us to use a single *text span* representation module for both mention detection and coreference resolution.

Following the previous work of Désoyer et al. (2015a), our preliminary experiments used mention-pair classification models, but their disappointing results prompted us to switch to an antecedent-scoring model.

As noted in section 4.2, the antecedent-scoring model is a notable improvement over mention pair classifiers, both in terms of empirical performance and in intuitive fitness to the task at hand, where the existence of a coreference link between a pair of mention is highly dependent on the other mentions in the same context (Denis and Baldridge 2007a, 2008), a fact that is hard to capture for a classification model.[6]

In line with most of the previous works on antecedent ranking, at inference time, given an antecedent score distribution for a mention, we simply greedily select the highest-scoring candidate as antecedent. This has proven to be good enough for a working system in practice and is less demanding in resources than more sophisticated alternatives.

Antecedent scoring

The antecedent scoring module uses a slightly modified version of the coarse-to-fine second-order inference mechanism from K. Lee et al. (2018). Instead of using a single antecedent score, it uses tow in a scaffold structure: a 'coarse"' score — determined by a function cheap to evaluate but with limited representation power — and a 'fine' score — more precise but more expensive to compute — used only for the antecedent candidates with the K highest coarse score, where K is an hyperparameter.

In our case, the coarse score $s_a^\flat(m, c)$ for a mention m and an antecedent candidate c is computed using a biased bilinear layer[7] as

$$s_a^\flat(m, c) = {}^{\mathrm{t}}g_m \mathrm{W}_\flat g_c + b_\flat \qquad [5.11]$$

where $\mathrm{W}_\flat$ and $b_\flat$ are learned parameters.

The choice of a biased bilinear function for this layer was mainly motivated by its computational efficiency (this operation is meant to be as fast as possible, since it will be applied to a large number of antecedent candidates) relative to its expressiveness (formally equivalent bilinear attention layers have proven their efficiency for a variety of tasks).

Let M′ be the set of the K antecedent candidates with the highest coarse score. For the elements of M′, we compute the fine score $s_a^\#(m, c)$ by applying a feedforward network to the representations

[6] On the other hand, using this approach, it is not clear how to predict the type of anaphoric link (as annotated in ANCOR). However, if e.g. Roesiger et al. (2018) shows promising results for this kind of predictions, this is still very much uncharted territory for coreference, where such granularity is still lacking in most annotated resources and evaluation metrics.

[7] Not to be confused with the *biaffine* layer introduced by Dozat and Manning (2017) and used for coreference resolution by R. Zhang et al. (2018)

of m and c and a feature vector $\phi(m, c)$ (see chapter 6 for details on these features). Formally

$$s_a^\sharp(m, c) = \text{FFNN}_a([g_m, g_a, \phi(m, c)]) \cdot \mathbb{1}_{c \in \text{M}'} \quad [5.12]$$

The total antecedent score for (m, c) is then the sum of the coarse and fine scores

$$s_a(m, c) = s_a^\flat(m, c) + s_a^\sharp(m, c) \quad [5.13]$$

which allows training both the coarse and the fine score for the top candidates.

Discourse-new score

While $s_a(m, c)$ represents the confidence that a given mention a is an antecedent of m, the discourse-new score $s_n(m)$ represents the confidence that m is the first mention of its chain, often improperly (Recasens 2010, p 24) called *anaphoricity score*. Since it is computed independently of the antecedent candidates' distribution, it depends only on the form of m, and an ideal model could use it to evaluate how well m fits the description of disourse-new mentions — for instance assigning a low s_n if m is a pronoun and a higher s_n if it is a definite description.

In practice, Vieira and Poesio (2000) findings suggest that even relatively local[8] syntax-based heuristics can be reasonably accurate for discourse-new detection and both shallow (Denis and Baldridge 2007b; Lassalle and Denis 2015) and deep (Wiseman et al. 2015) learning approaches were found to benefit from learning an explicit 'anaphoricity' score.

K. Lee et al. (2017) chose to fix this score to 0, in order to '[remove] a spurious degree of freedom'. This does not reduce the expressivity of the model, which simple has to learn to affect negative antecedent scores to all antecedent candidates for mentions that resemble discourse-new mentions. However, this puts all the burden of identifying the common intrinsic characteristics of discourse-new mentions on the $s_a(m, c)$ score, thus spending part of the expressiveness of the s_a function for things that do not depend on c.

In addition to this, since singleton mentions are always discourse-new, and given the high ratio of singletons entities, discourse-new mention are by far the most common mentions in our data (62.55 % in the training set for instance). Therefore, leaving the detection of these mentions solely to the antecedent scoring module would make the majority class heuristic (in this case, always assigning a negative antecedent score to all the candidates) very easy to fall in.

Given these points, we elected not to use a constant discourse-new score, and to train a separate scoring layer

$$s_n(m) = \text{FFNN}_n(g_m) \quad [5.14]$$

Mention embeddings refining

The antecedent-scoring procedure described so far now is purely *local* (Denis and Baldridge (2007b, 2008) and McCallum and Wellner (2004), see section 4.2 for details), since coreference

[8] Uryupina (2003) and Poesio et al. (2005) show even more encouraging results, but use non-local features.

chains are built greedily, with no possibility of global supervision. As mentioned earlier, in practice, this heuristic works reasonably well,[9] but it can be too simplistic in the general case.

In that regard, the mention refining procedure proposed by K. Lee et al. (2018) is as a step in the direction of global coreference resolution that has the advantage of being both relatively cheap (as opposed to more involved earlier methods such as ILP decoding) while still being reasonably efficient.[10] It consists in modifying (*refining*) the initial representation of spans by combining them with the representations of their antecedent candidates, using the aforementioned coarse score $s_a^\flat$ as a weighting function.

Formally, the refined representation of m is given by

$$g'_m = \mathrm{R}\left(g_m, \sum_{c \in \mathrm{M}} s_a^\flat(m,c) g_c\right) \tag{5.15}$$

where R is an elementwise multiplicative sigmoid gate (Cho et al. 2014; Hochreiter and Schmidhuber 1997) and

$$\begin{aligned} \lambda(x,y) &= \sigma(\mathrm{W}_\mathrm{R}[x,y] + \beta_\mathrm{R}) \\ \mathrm{R}(x,y) &= (1-\lambda(x,y))x + \lambda(x,y)y \end{aligned} \tag{5.16}$$

Beyond the obvious parallel with the gating mechanisms in gated recurrent layers, this is also reminiscent of the method used to combine heterogeneous representations used in Balazs and Matsuo (2019) and Miyamoto and Cho (2016) (to combine word- and character-level representations of tokens in these cases).

Equation 5.12 then becomes

$$s_a^\sharp(m,c) = \mathrm{FFNN}_a([g'_m, g_a, \phi(m,c)]) \cdot \mathbb{1}_{c \in \mathrm{M}'} \tag{5.17}$$

Note that since we do not proceed at the document level, refining only makes sense for the representation of m and not for its antecedent candidates (as we do not have access to *their* antecedent candidates at that time). Consequently, it would not make a lot of sense to reiterate the refining procedure (K. Lee et al. (2018) report a maximum return on investment for two consecutive refinements), since only the representation of m could change between the refining rounds.

5.3 Training

The inference algorithm presented in fig. 5.1 process a whole document in a single pass, iteratively detecting mentions and finding their antecedents, thus solving the coreference resolution task in an end-to-end setting. As shown by K. Lee et al. (2017), this kind of end-to-end processing can also be applied to the task of training a coreference resolution system. However, end-to-end

[9] Possibly because the repartition of complexity of coreference relation detection is heavily skewed, with a large proportion of naturally occurring coreferences being easily detected with simple heuristics (see chapter 2).

[10] Regrettably, their reported experiments did not include an evaluation of the benefits of this technique independently of the use of contextual embeddings —whose impressive gains overshadow their other interesting contributions— but its positive effect is still clear.

training comes with a high cost, both in terms of computations and memory usage. In our case, the separation of end-to-end coreference resolution in two components gives us a natural way to train our system at a lower cost: we can train mention detection and antecedent scoring independently, while still sharing a common span representation module.

5.3.1 Optimization algorithm

While second-order methods (Martens 2016) and evolutionary strategies (Gaier and Ha 2019) have been suggested to be more relevant in some contexts, at the time of writing, the overwhelming majority of neural network optimizations use variants of the Stochastic Gradient Descend (SGD) algorithm instead (LeCun et al. 2012). As with other hyperparameters, choosing among these variants can be a daunting task, with no clear answer on which one is the best suited for a given architecture. Exhaustive comparison is also far too costly to consider, particularly since most of these variants require further hyperparameter tuning to reach their full potential. In this context, Adam (Kingma and J. Ba 2014) has the advantage of having been empirically shown to be efficient for many neural network architectures and natural language processing tasks (notable examples being Devlin et al. (2018) and K. Lee et al. (2017)). It also requires very little tuning of hyperparameters (except for a learning rate schedule, see section 5.3.2). Recent advances (Dozat 2016; L. Liu et al. 2019; Reddi et al. 2018) have suggested that further modifications to Adam might make it even more efficient, but these have not yet received enough third-party attention to consider them clear successors.

Considering these points, for both mention detection and antecedent scoring, we optimize the trainable parameters of the architectures described in section 5.2.1 using the standard version of Adam from Kingma and J. Ba (2014), with a single modification: the weights of the embedding layers are only updated when their current gradient is non-null, instead of drifting with their current momentum (the so-called 'sparse' or 'lazy' version Adam used in PyTorch for sparse layers). Loss gradients with respect to the weights are computed by backpropagation (LeCun 1988; Rumelhart et al. 1986) using the automatic differentiation facilities of the Pytorch library (Paszke et al. 2017), see sections 5.3.4 and 5.3.5 for details on the loss functions considered.

5.3.2 Learning rate

Even when using adaptative optimization algorithms such as Adam —where the learning rate is dynamically adapted to the topology of the parameters search space—, designed to relieve practitioners from the hassle of fine-tuning yet another hyperparameter, the base learning rate still has a significant influence on the training process.[11] Recent works (Loshchilov and Hutter 2019) go further and recommend the use of Adam with an explicit learning rate schedule (see also Devlin et al. (2018) for a notorious example). L. Liu et al. (2019) brings some theoretical justifications for this, in particular for the ubiquitous warmup strategy.

Given the lack of definitive answers regarding the best way to choose a learning rate, we experimented with some popular choices. In table 5.8, we compare the results of learning mention detection with a constant base learning rate for a few common values and with a

[11] See for instance Karpathy and Johnson (2019).

common schedule: a geometric step decay with warmup. These results are of course not to be taken as a claim of exhaustivity, and further theoretical work in this area is still very much needed.

5.3.3 Regularization

Among machine learning systems, neural networks occupy a special space with regard to the problem of overfitting: on one hand, the ratios between their number of parameters and the size of the datasets they are usually trained on are usually much higher than those of other systems, but on the other hand this does not seem to harm their generalization capacities.[12] In fact, it seems that on the contrary, beyond a certain number of parameters, the classical U-shaped bias/variance risk curve starts decreasing again and adding more parameters yields architectures with more generalization power (Belkin et al. 2018). This has also been observed recently in experiments with Transformer-based language models (Devlin et al. 2018; Y. Liu et al. 2019; Radford et al. 2019; Shoeybi et al. 2019). Given this, the appropriateness of regularization for neural network architectures might be questioned. Indeed, findings from C. Zhang et al. (2016) suggest that:

> *Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.* (C. Zhang et al. 2016)

However, it *may* still improve generalization performance and several such methods have been experimentally shown to not only do this, but also to stabilize and to speed up the learning process (see for instance J. L. Ba et al. (2016) and Ioffe and Szegedy (2015)). Given this, it seems sensible to at least try to apply some regularization to our base architecture, even if we could not afford an exhaustive exploration of the whole landscape of regularization strategies. We experiment with three well-known regularization techniques: dropout, weight decay and layer normalization.

Dropout

Dropout, first proposed by Hinton et al. (2012) is a regularization method that reduces the tendency of neural networks to overfit during training by randomly zeroing a certain amount of coordinates from the outputs of the input and hidden layers. Intuitively, this makes it harder for a system to remember a strict mapping from the training samples to the corresponding gold output (which would lead to very poor generalization) and forces it to learn more general structures. It can also be seen as an approximation of model ensemble averaging (Warde-Farley et al. 2013).

Several variations on dropout have been proposed, most notably DropConnect (Wan et al. 2013) which zeroes weights instead of activations ; Gaussian (Srivastava et al. 2014) and variational (Kingma et al. 2015) dropouts, which fuzz the activations with random noise instead of zeroing them ; recurrent dropout (Gal and Ghahramani 2015) and Zoneout (Krueger et al. 2016), where a single dropout mask is applied to each time step of a recurrent layer ; and curriculum dropout (Morerio et al. 2017), where the dropout rate p is gradually increased during training to accelerate

[12] For a comprehensive review on this, see Weng (2019)

the learning process in early epochs while still preventing overfitting in the long run. However, while all of these variations reportedly improve the performances of some neural network architectures on specific benchmarks, there is yet no consensus on their benefits in general and their support in off-the-shelf neural network frameworks is still lacking.

For this reason, we limit our study to the original Bernoulli dropout as formulated by Srivastava et al. (2014). In this formulation, dropout is only applied during training and the activations are also scaled up by a factor of $\frac{1}{1-p}$ in order to prevent a spurious boost of the mean of their distribution at test time, which could hamper the operation in the next layer.

We apply dropout on the inputs of all the neural layers in our architecture, with the rates reported in section 5.4.1.

Weight decay

Introduced by Hinton (1987) and formalized by Krogh and Hertz (1991), weight decay, like dropout, attempts to reduce overfitting by reducing the long-term influence of the specific noise of the training samples on the network weights through the introduction of a regularization term at each weight update. In the formulation of Hanson and Pratt (1988), shown by Loshchilov and Hutter (2019) to be more suitable in the general case than Krogh and Hertz (1991)'s,

$$\theta_{t+1} = (1 - \lambda)\theta_t - \alpha_t \nabla f_t(\theta_t) \qquad [5.18]$$

where θ_t, α_t and $\nabla f_t(\theta_t)$ are respectively the weight vector, the learning rate and the error gradient at training step t and $\lambda \in [0, 1[$ is the weight decay rate.

In table 5.9 we report the results of using weight decay with different decay rates, using the AdamW variant of Adam (Loshchilov and Hutter 2019), which gives a true weight decay instead of the L^2 regularization common in older implementations.

Layer normalization

Layer normalization, first proposed J. L. Ba et al. (2016), is an improvement over older techniques such as batch normalization (Ioffe and Szegedy 2015) and weight normalization (Salimans and Kingma 2016). All of these techniques consist in *normalizing* the output of a neural layer, by shifting and scaling the resulting activations to make their empirical mean and standard deviation constant across the training dataset. Formally, given the output $x \in \mathbb{R}^d$ of a neural layer, the corresponding layer-normalised output x' is

$$x' = \gamma \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \qquad [5.19]$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ are respectively the empirical mean and the standard deviation of x, seen as an i.i.d. sample of some distribution, $\varepsilon \in \mathbb{R}^+$ is a small positive number (typically 1×10^{-5}) meant for numerical stability and $\gamma \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ are learnable elementwise affine transform parameters.

This normalization aims at reducing *internal covariate shift* (Ioffe and Szegedy 2015). It is motivated by the observation that while training deep neural networks, the last layers are

significantly harder to train due to the fact that the distribution of their inputs (the outputs of previous layers) tend to vary significantly during training. This bears some similarities with the phenomenon known as *covariate shift* (Shimodaira 2000) where variations in the input distribution of a machine-learning system hampers the efficiency of the learning process.

The real impact of activation normalization on internal covariate shift is currently challenged. For instance, Santurkar et al. (2018) dispute both the claims that internal covariate shift hampers training and that batch normalization reduces it. The authors instead ascribe the improvements observed when using in batch normalization to its smoothing effect on the loss function as observed by the normalized layer.

However, empirical results tend to show that these normalization techniques have a positive impact on both learning speed and validation performance even for tasks and architectures not studied in the original works where these techniques were first introduced. Specifically, layer normalization, for its ease of use at both training and test time is at the core of the Transformer model (Vaswani et al. 2017) where it is used in the add-and-norm residual connection as a mechanism to regulate gradient explosion (Chen et al. 2018).

In our case, layer normalization is applied on the output of sequence to vector encoders, ie after the FFNN_c, FFNN_o and π_s layers.

5.3.4 Mention detection

Since we cast mention detection as a classification problem, training the corresponding architecture is relatively straightforward, as neural classification is now a very common problem. Accordingly, we train the parameters of N_g and d end-to-end, by minimizing the negative log-likelihood of the correct class in the $softmax$-normalized score distribution given by d (or, equivalently, the cross-entropy of this score distribution with respect to the corresponding punctual distribution). Formally, given the class score distribution $v = \text{FFNN}_d(g_e) \in \mathbb{R}^c$ for a span e for which the correct class is $y \in [\![0, c]\!]$, this loss is

$$\mathscr{L}_d(v, y) = -\log(\text{softmax}(v)_y) \qquad [5.20]$$

which is simply averaged over individual samples in order to get a batch-level loss.

Overcoming class imbalance

The biggest obstacle in considering mention detection as a classification task is the major[13] imbalance between mention and non-mention spans, making the majority class heuristic far more tempting for the system than in more balanced classification tasks. Moreover, even with the steps taken to reduce the complexity described in section 5.2.3, the size of the training set resulting from considering all spans is still considerable.

A natural way to solve both issues is to reduce the size of the training dataset by removing a certain number of non-mention spans from it, a procedure known as *undersampling* (Haixiang et al. 2017). According to Buda et al. (2018), undersampling does not necessarily have a positive

[13] Only about 1.81 % of all considered spans are mentions in the development dataset for instance.

effect on learning (although the authors only considered *balanced* undersampling, which brings the minority and majority class to the same ratio) but it does not usually have a negative effect. Furthermore, the speed gain in our case was critical. We report the results of experimenting with various sampling ratios in table 5.10.

5.3.5 Antecedent scoring

Following K. Lee et al. (2017, 2018) [14] , we train our antecedent scoring module by minimizing the sum of the log-likelihood of all the correct antecedents for each mention.

Formally, given a mention m with antecedent candidates $c_1, ..., c_k$, $\mathrm{Y} = \{i \in [\![1, n]\!] \,|\, c_i \text{ is an antecedent of } m\}$ and $\mathrm{A} = [s_n(m), s_a(m, c_1), ..., s_a(m, c_k)]$, the per-sample loss is

$$\mathscr{L}_a(\mathrm{A}, \mathrm{Y}) = -\log \sum_{y \in \mathrm{Y}} \mathrm{softmax}(\mathrm{A})_y \qquad [5.21]$$

as with mention detection, the batch-level loss is simply the (unweighted) average of all per-sample loss.

Due to our *local* resolution of coreference, we could not directly use the differentiable relaxed coreference metrics introduced by P. Le and Titov (2017) or the reinforcement learning approach of Clark and Manning (2016a) since they requir document-level processing.

As noted in chapter 4, some works on coreference resolution using machine learning (and all of those using *neural* architectures) train their coreference resolution modules using predicted mentions. In our setting, this pose a number of issues, both from a technical and a methodological standpoint.

Firstly, if we did train our system using system mentions, we would have to choose a target antecedent distribution for spurious mentions resulting from precision errors. It is clear that this target distribution could not include any mention, leaving us with the choice of marking them as discourse-new, which they are not (contrarily to the no-singleton settings where spurious mentions can just be categorized as singletons to be discarded), leading our system to learn wrong outputs. Another option would be to consider only gold mentions but to allow system mention as antecedent candidates. Still, this would considerably complicate the training process and it is not clear that it would meaningfully improve the learning procedure.

Secondly, the best way to deal with recall errors — gold mentions missing in the system output — is also not clear: including them would not be consistent with the proposed ways to deal with precision errors, but ignoring them would remove precious from the already limited training dataset.

Lastly, even if we did devise a relevant scheme for introducing spurious mentions in the antecedent-finding learning phase, it would make the intrinsic performances of this module far harder to analyse, as they would in this case be dependent on the performances of the mention detection module.

[14] For this, we chose to trust their findings that using e.g. the previously popular weighted slack-rescaled margin loss (Wiseman et al. 2016, 2015) did not help training. This is also partly motivated by the need to tune yet more hyperparameters (the error cost weights) to use this family of loss.

For all these reasons, following the example of Bengtson and Roth (2008), Culotta et al. (2007) and Rahman and Ng (2009) among others and unlike K. Lee et al. (2017) [15] , we train this module exclusively on gold mentions [16] .

Finally, in order to reduce the complexity of the training procedure, we only consider antecedent candidates up to the 100 previous mentions, with the effect that coreference chains with mentions further apart than this limit are effectively cut [17] . This can seem to be limiting, but we argue that this simplification is reasonable. From a purely technical point of view, this concerns less than 7 % of all mentions while vastly simplifying our operations, given that the documents we consider can include several thousands of mentions. This is also not inherently limiting at inference time, since it can be easily disabled.

Moreover, from a theoretical point of view, in the discourse model framework that we adopt, it is not clear that the relation between mentions that are very far apart are coreferent *stricto sensu*, as it could be argued that they do not in fact refer to a single persistent entity in the discourse space, but rather that the identity relation holds between distinct entities existing in disjunct segments of the discourse (see chapter 2). That said, these matters fall well out of the scope of the present work, and in the current state of our conventions for the evaluation of coreference resolution, limiting the scope of antecedent-finding should be seen as a purely technical simplification, which is in any case *not* reflected in the experimental results that we report are of course given against the actual gold test clusters.

5.3.6 Spans representations

The span representation module N_g is the centrepiece of our whole architecture and condition most of its performances on both subtasks. As such, it is of critical importance to ensure that the representations it provides to the rest of our system are as good as possible for these tasks.

It is a well-known fact that pre-training input representations on auxiliary, possibly simpler, tasks is highly beneficial for subsequent, more involved, tasks (Devlin et al. (2018), Mikolov et al. (2013) and Peters et al. (2018) are possibly the most notorious examples of such pretraining in recent years, while Ruder (2019) provides a recent and comprehensive review). Since our architecture already consists of two tasks and one of them —mention detection— is intuitively simpler, it seemed natural to try to reuse the parameters learned in mention detection to initialize N_g for antecedent scoring.

However, while this *sequential* transfer learning procedure can be beneficial, it can also fail due to the tendency of neural networks to *catastrophically forget* previous training (French 1999) in some contexts.

[15] Although this characterization is somewhat misleading, since they do not actually train on mentions, gold or system, but on *all spans*.

[16] Note, however, that this only concerns mentions boundaries, as we do not use any other gold features. In particular, unlike Bengtson and Roth (2008), we do not use gold mention types.

[17] Note that this is fundamentally different from the hard antecedent candidates limit set by K. Lee et al. (2017) and lifted by K. Lee et al. (2018), as theirs was a mean to work with the high number of candidate *spans*. In their case, the reduction of number of candidate *mentions* resulted from the truncation of documents to 50 sentences.

It is still not yet clear what kind of pretraining is beneficial to subsequent tasks and what kind simply amounts to another random initialization, or worse, yields input representations that are so unsuited for the subsequent task that it makes training the corresponding neural network harder than random initialization. Therefore, to assess the applicability of sequential transfer learning, we experiment with both the mention detection → antecedent scoring order and the reverse antecedent scoring → mention detection. This allows us to report results for independent training (by taking the result of the first step in both cases) and with both kinds of pretraining (from the result of the second step)

In addition to sequential transfer learning, we also experiment with a variant of the hard parameter sharing multi-task learning procedure described by Ruder (2017) and Søgaard and Goldberg (2016) and used with success by Sanh et al. (2018) for several NLP tasks including coreference detecion (in the CoNLL-2012 shared task setting). It consists in training two or more models that share some parameters in parallel by randomly sampling a model at each training step, train on it on a single batch of the corresponding data and repeating this until convergence. This is an easy way of training models that share some of their parts on different tasks with almost no implementation costs and without requiring multi-task objectives, which are notoriously hard to design properly.

In our case, the models are the coreference resolution and the antecedent scoring one and they share the span embedding module. In this variant, instead of randomly sampling between the models, we simply cycle through them, training for N steps on mention detection, then M steps on antecedent scoring, evaluating each model on the development dataset after each epoch. As in the sequential case, we stop training as soon as one of the models development performances did not improve in the last p epochs.

5.4 Experiments

To assess the adequacy of our architecture, we train and evaluate it on ANCOR, using the train/development/test split from Grobol (2019). The models for both mention detection and antecedent scoring are trained end-to-end as described in section 5.3 until their performance on the development data cease to improve, with an arbitrary patience of 5 epochs, which seems to be a good compromise between giving the procedure unrestricted freedom to explore the parameters space and limiting our computation budget. The performances that we report are those of the dev-best model, both on development and test data.

The mention detection metrics are the usual precision, recall and F_1 score (see section 2.4.1). For coreference, we report the CoNLL coreference metrics MUC, B^3, $CEAF_e$ and MELA (*alias* CoNLL), and BLANC (see section 2.4.2) on the system coreference clusters including singlerons. These metrics are computed using scorch, as described in section 2.4.3. In particular, corpus-level scores are computed by micro-averaging for all metrics using the number of mentions per document as weights instead of using ad-hoc averaging schemes for MUC and BLANC, in order to reach a better consistency across metrics and prevent an overly large influence of the largest documents — which is particularly relevant for ANCOR, given the heterogeneity of the sizes of its documents.

5.4.1 Hyperparameters

Since our goal was to assess the viability of this architecture for the particular case of spoken French more than pushing the boundaries of quantitative evaluations,[18] the hyperparameters of both our architecture and its training setting were not extensively or systematically fine-tuned (however, see section 5.4.1 for experimental results on *some* hyperparameters), nor did we experiment with model ensembling or boosting methods.

Input embeddings

Unless otherwise specified, all input embeddings are initialized randomly by sampling from the $\mathcal{N}(0, 1)$ distribution and tuned during all the training phase.

The word embeddings e_w are initialized using the FastText embeddings pretrained by (Grave et al. 2018) on a mixture of Wikipedia and Common Crawl data. The words for which no vector exists in these pretrained embeddings are kept (mostly to avoid discarding oral-specific tokens) with their default initializations. This default initialization is also used for the special tokens `<start>`, `<end>` (marking the beginnings and ends of utterances) and `<unk>` (for out-of-vocabulary words).

To take into account the casing information in the least disruptive way, when digitizing a word, we first look for an embedding with the same casing and fallback successively to looking for an embedding for its lowercasing and using the `<unk>` embedding. This is still not completely ideal, as it generate for instance two distinct embeddings for a word if it occurs more than twice at the beginning of a sentence, but it also let us keep distinct embeddings for proper noun homographs (such as *paris* vs. *Paris*) while requiring no systematic lexicon.

The character lexicon consists of all the characters encountered in the training dataset, with the addition of an `<unk>` character. The special tokens `<start>` are `<end>` treated as consisting of the single character `<no>`.

The dimensions and vocabulary sizes of all the input embeddings described in this chapter are reported in table 5.1, including those for the span and pair features (see)chapter 6).

Table 5.1: Embedding dimensions in DeCOFre baseline

Feature	Symbol	Vocabulary size	Dimension
Word	e_w	10 191	300
Character	e_c	83	50
Span length	ϕ	10	20
Words distance	ψ	10	20
Utterances distance	ψ	10	20
Mentions distance	ψ	10	20
Speaker agreement	ψ	2	2
Overlap	ψ	2	2

[18] Especially since none exists that we can meaningfully compare to.

Size restrictions

As mentioned in section 5.2.3, we only consider spans of at most 25 words to reduce the time and material requirements. This value is chosen arbitrarily to be superior to the length of the longest mention in ANCOR (17 words). The span contexts considered are of size 10 on both sides.

For antecedent-finding, as mentioned in section 5.3.5, only the 100 previous mentions are considered for antecedent scoring, and only the K = 25 best-scoring antecedents are kept for fine-scoring.

Neural layers

Table 5.2 details the parameters of the feedforward layers used in our architecture. They all use at most one hidden layer, as in our experiments, increasing their depth did not seem to improve their usefulness, while it significantly increased the training time. This is not very surprising, since the global architecture in itself is already considerably deep. Some of them exhibit unusually large input sizes to deal with the fact that their inputs are concatenations of representations from several sources.

Our default non-linearity is the leaky rectifier[19]

$$\left|\begin{array}{rccl} \text{LReLU}: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \text{LReLU}(x) = \max(0,x) + 0.01\min(x,0) \end{array}\right. \quad [5.22]$$

a slightly modified version of the rectifier non-linearity (Nair and Hinton 2010) that does not suffer from the "dying ReLU" problem and generally performs well for many natural language processing tasks (Eger et al. 2018).

Table 5.2: Feedfoward networks parameters in DeCOFre baseline

Layer	Layers/Dimensions	Nonlinearities	Dropout
FFNN_c	(300, 50)	None (Layer Norm)	0.6
FFNN^i_α	(950, 300, 1)	LReLU, None	0.2, 0.2
FFNN_o	(4400, 500, 500)	LReLU, LReLU (Layer Norm)	0.6, 0.2
π_s	(520, 500)	LReLU (Layer Norm)	None
FFNN_d	(500, 200, 3)	LReLU, None	0.6, 0.6
FFNN_a	(1064, 150, 1)	LReLU, None	0.4, 0.4
FFNN_n	(500, 150, 1)	LReLU, None	0.4, 0.4

- *Layers/Dimension* lists the dimensions of *all* layers, including input and output.
- *Nonlinearities* and *Dropout* rates are given for forward connection, with the convention that dropout and non-linearity are applied respectively before and after the corresponding connection.

[19]The notation (L)ReLU comes from the conflation of the name of activation —the (leaky) rectifier— and the designation of the neural units of which it is the non-linearity —(**L**eaky) **Re**ctified **L**inear **U**nits.

Table 5.3 lists the hyperparameters for the recurrent layers, both of which are bidirectional (hence the 2× term for the dimension) and use the default Pytorch v1.4.0 bindings for the CuDNN library (Chetlur et al. 2014).

Table 5.3: Recurrent layers parameters in DeCOFre baseline

Layer	Layers	Hidden dimension
LSTM_w	2	2×300
GRU_c	1	2×150

5.4.2 Baseline results

Mention detection

Table 5.4: Mention detection evaluation (ANCOR test)

System	P	R	F
ANCOR			
Grobol et al. (2017b)	57.28	77.07	65.72
Grobol (2019)	82.99	89.07	85.87
Our baseline (on test)	86.83	88.41	87.62
Other corpora			
Uryupina (2010) (SemEval)	79.8	76.4	78.1
Soraluze et al. (2017) (EPEC)	74.67	74.47	74.57
Poesio et al. (2018) (ARRAU)	79.33	86.16	82.60

Table 5.4 presents the results of our experiments with mention detection compared to the baseline of Grobol et al. (2017b) — which consists in merely extracting all the NP from the output of an off-the-shelf parser — and those reported in Grobol (2019) for an earlier state of DeCOFre. To give some sense of the performances that could be hoped for, given the lack of relevant work on ANCOR, we also include the performance of other mention detection systems evaluated on comparable corpora for other languages, but *these results should of course not be taken as fully comparable with ours.*

Coreference resolution

Table 5.5 presents the performances of our system for coreference resolution and compare it with those of previous works. Note that we did not compare with the performances of systems designed for the CoNLL-2012 setting (and in particular those of K. Lee et al. (2017, 2018)) since there is in general no simple way to provide them with gold mentions[20] at either training or

[20]In the usual meaning used in this work and not in the 'anaphoric gold mentions' sense used in K. Lee et al. (2017).

Table 5.5: Coreference resolution evaluation (ANCOR test)

	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
Désoyer et al. (2015a)	—	—	63.5	—	—	83.8	—	—	79.0	75.3	—	—	67.4
Grobol (2019)	72.3	47.7	57.3	89.7	71.0	79.2	72.8	86.0	79.4	72.0	78.2	60.1	65.7
Our baseline	73.71	54.22	62.14	89.51	75.29	81.67	77.28	88.15	82.26	75.36	80.32	66.71	71.12

test time, nor to make them distinguish between singleton mention and non-mention spans without significantly modifying them.

As mentioned in chapter 4, the existing works on ANCOR have been developed in different paradigms and are not easily comparable to ours. This is particularly true for Désoyer et al. (2015a), which relies on gold features —including mention anaphoricity—, and as such was able to get high scores on all metrics with a relatively simple system. Their results should thus be considered as an assessment of the usefulness of these features rather than a realistic benchmark, especially since it only reports F_1 scores and used an undocumented dataset split, which limits the interpretability of their results

With this reserve, the results presented in table 5.5 prove that is possible to design an accurate coreference resolution system for spoken French without relying on specific resources or extensive linguistic knowledge —and indeed obtain better quantitative results than a system using gold features.

5.4.3 Transfer learning

As mentioned in section 5.3.6, there are several alternatives while training an architecture with shared parameters. We report here a comparison between the performances of our baseline model in some training configuration for both tasks: without pretraining, with pretraining on the other tasks and with quasi-simultaneous training by alternating training on 2 batches for mention detection and 1 batch for antecedent detection. Apart from this, all the settings of these experiments are the same as those of our baseline, whose configurations are marked with an asterisk (*).

Table 5.6: Influence of training setting on mention detection

ANCOR dev			
Setting	**P**	**R**	**F**
*No pretraining	86.89	88.15	87.52
Pretraining	88.62	84.22	86.36
Simultaneous	87.23	86.47	86.85

Table 5.6 compares the results between the three methods for mention detection, with a clear improvement for the case where the span embeddings were not pretrained using antecedent scoring. Simultaneous training does not seem to help either, which suggests that training on antecedent scoring actively degrades the relevance of the span embeddings for mention detection. These results are quite intuitive: during each epoch antecedent scoring training, a small number of spans (the gold mentions) are each seen by the network around 100 times (the maximum number of antecedents), with the consequence that the resulting spans representations are highly optimized for this specific category of spans.

Table 5.7: Influence of the training setting on coreference resolution

	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
No pretraining	75.02	54.61	62.8	90.52	75.12	81.99	77.17	88.45	82.33	75.71	82.03	65.18	70.15
*Pretraining	73.71	54.22	62.14	89.51	75.29	81.67	77.28	88.15	82.26	75.36	80.32	66.71	71.12
Simultaneous	71.93	48.54	57.65	89.4	72.88	80.23	73.81	87.79	80.12	72.67	77.51	62.68	67.18

Table 5.7 compares the results for coreference resolution, with a much less clear situation. Here, the results are slightly better without pretraining for the CoNLL metrics and slightly worse for BLANC, due to the more balanced results when using pretraining. In the end, our choice for our baseline was a pretrained model, BLANC being more relevant for our data.

What is clear, however is the situation of the simultaneous training setting: for coreference resolution, too, it noticeably degrades the performances. In addition to these experiments, we tried some other simultaneous training configurations, with different task repartitions, with results that were similar to or worse than these, with significantly lower learning speed (which was to be expected) and a tendency to diverge in later training stages.

5.4.4 Hyperparameters influence

In addition to the baseline results, we report here the results of some of our experiments on choosing hyperparameters. Most of these experiments were done on only on the mention detection step of our model, due to the cost of training the antecedent scoring model. The settings used for our baseline results are marked with an asterisk (*).

In table 5.8, we study the influence of the base learning rate on the quality of the mention detection model by testing some common values with a constant learning rate and the default Adam base rate of 1×10^{-3} with a warmup over the first 1000 batches and a multiplicative step decay of 0.7 per epoch, which corresponds to an average learning rate of approximately 3×10^{-4} (the best constant rate in our experiments over the 10 first epochs, our upper limit for coreference resolution).

Consistently with the current understanding of the influence of learning rate for Adam (Karpathy and Johnson 2019), 3×10^{-4} seems to be the safest alternative between the often too high 1×10^{-3}, which leads to unstable convergence on the training set and the too low 1×10^{-4}, which leads

Table 5.8: Influence of the learning rate schedule on mention detection

ANCOR dev				
Schedule	**Base LR**	**P**	**R**	**F**
Constant	1×10^{-4}	84.58	87.23	85.88
Constant	3×10^{-4}	84.82	89.94	87.06
Constant	1×10^{-3}	83.66	89.52	86.49
*Step decay (0.7)	1×10^{-3}	86.89	88.15	87.52

Table 5.9: Influence of weight decay on mention detection performances

ANCOR dev			
Decay rate	**P**	**R**	**F**
1×10^{-5}	87.27	83.65	85.42
*0	86.89	88.15	87.52

to a slow training process and degraded generalization capabilities, possibly due to overfitting.

However, it seems that decaying the learning rate during the training process is a better alternative. A tentative interpretation could be that a high learning rate in the first epoch allows a rapid exploration of the parameters space to find a promising region, while the lower values in the later epochs allow for more precise tuning within this region. That said, this is a very small sample of the possible choices for learning rate schedules, and many more experiments would be needed to confirm these findings and determine an optimal choice.

In table 5.9, we report the outcome of using weight decay regularization to train the mention detection module. Even with a relatively mild weight decay rate (1×10^{-5} for a base learning rate of 1×10^{-3}), the negative impact on the performances is significant.

This might be a consequence of the sparsity of the training signal for mention (due to the aforementioned heavy imbalance between spans classes), our use of other regularization methods (Dropout and Layer Normalization) which already makes the learning process harder (if hopefully more robust in generalization), a combination of both or unrelated factors. In any case, this result was sufficiently bad to discourage us from further experiments in this direction.

5.5 Conclusion

In this chapter, we proposed an end-to-end coreference resolution system for oral French, taking into account the peculiarity of this domain (lack of dedicated resources, non-standard and irregular language), the specificities of the available data (document nature, size and characteristics of mentions and coreference chains) and our material resources. Our architecture

Table 5.10: Influence of undersampling rate on mention detection perfomances

ANCOR dev			
System	**P**	**R**	**F**
90 %	81.50	90.12	85.60
*95 %	86.89	88.15	87.52

is an hybrid of recent end-to-end neural architectures — which gives us a relative independence from linguistic resources — and more classical pipelines — which allows us to use a less constrained definition of entities and significantly reduces the computational cost of training our system.

The quantitative results obtained by this system are on par or better than those of previous attempts on the same data, which proves that our design is at least sensible for this setting. However, there seems yet to be a large room for further improvements.

The success of deep neural networks in natural language processing are still relatively recent, and in many areas, the reasons for their efficiency are still unclear. There is also a clear lack of definitive answers on which techniques are optimal for a given task. In this work, we have explored but a small fraction of the alternatives the current state of the art proposes. In further works, a much more thorough review of our choices should be conducted to identify the parts of our architecture that are responsible for its good results, those that are responsible for its mistakes and ways to improve it, drawing inspiration from C. Zhang et al. (2019) for instance. To this effect, we have tried to suggest interesting alternatives to our choices along this chapter, but given the breadth of the field and its current rate of expansion, many others are bound to surface in the years to come, either from human design or automatic architecture search (Wong et al. 2018; Zoph and Quoc V. Le 2017).

Even with our current architecture, the size of the hyperparameters space is consequent — both for the neural networks themselves and their training process — and our explorations were limited by material and temporal constraints to the relatively safe region of common values and did not range very far even there. Further work should thus also be concerned by making more thorough assessments of the influence of these choice and systematically improving them (Bergstra et al. 2011). Real-world usage of our architecture could also benefit from pragmatic performance scraping techniques, such as boosting or model ensembling, which proved valuable for comparable works (K. Lee et al. 2017).

Finally, in line with the advantages we found in using a pipeline instead of a fully end-to-end system, it is likely that there are benefits to be reaped from reintroducing explicit knowledge in coreference resolution systems using techniques that have received little attention in the recent years compared to the rush towards improving neural architectures. For instance global decoding techniques, such as Integer Linear Programming (Denis and Baldridge 2007b, 2009) or

entity-mention models (which have seen some attention from Wiseman et al. (2016) and some recent exploratory work from Fei Liu et al. (2019)) could prove beneficial.

Regarding mention detection, given its inherently syntactic nature, formulations closer to those of successful syntactic parsing systems will probably be both more legitimate and more efficient, given the very encouraging results in that regard from B. Wang et al. (2018). Orthogonally to the question of integrating human knowledge of coreference at the architectural level, we also give in chapter 6 some propositions of augmentations using explicit features from external linguistic knowledge.

Chapter 6

Explicit knowledge augmentations for coreference resolution systems

In chapter 5, we have shown that it is possible to develop a coreference resolution system for spoken French using almost no *a priori* knowledge, by relying instead on dense vector representations of inputs learned end-to-end by a neural network architecture. In this chapter, we present the results of our investigations on techniques to introduce explicit knowledge in this neural architecture and the benefits that such knowledge could bring to a modern coreference resolution system. Here, by 'explicit knowledge' we mean information that is not directly available using only coreference annotated resources, in which we include not only knowledge coming from external resources and preprocessing tools, but also information that comes from the application of *a priori* linguistic knowledge —e.g. a linguistic model of reference[1] — to the raw data in a coreference corpus.

For this, we face two main issues: how to include such knowledge in a pre-existing neural architecture, and how to cope with the lack of reliable knowledge and preprocessing tools for spoken French, noted in section 5.1 and which led us to develop a knowledge-poor system in the first place. These issues are not necessarily independent: since the knowledge that we would like to incorporate in our system is not completely reliable, it is critical that it is included in such way that our system has the opportunity to learn during its training when and how this knowledge can be depended upon and when to ignore it.

In section 6.1, we study raw features —features that do not depend on external knowledge, including some that are already included in our baseline system, but that we did not describe in details in chapter 5.

In section 6.2, we experiment with weakly supervised knowledge —that depends on the availability of unannotated linguistic material in the target language.

Then in section 6.3, we study the role of shallow linguistic knowledge, by adding general-purpose word-level linguistic information to the inputs of our system.

[1] Even though the work presented here stays at a lower level of sophistication.

Finally, in section 6.4, we work on the addition of structural linguistic knowledge —such as syntactic parsing or named entity information— that are directly linked to aspects of coreference resolution.

6.1 Raw features

We call *raw features*, all features that do not depend on resources beyond the available content in the raw source material. Given that our work is about speech transcriptions, this does not mean that this content is entirely void of annotations, since —as we noted in section 3.2.2— accurate representation of spoken language in written transcription implies the presence of information beyond the words uttered. Accordingly, we also consider features that depend on speaker identification, utterances boundaries, etc. information as raw, in the same sense that features relying on punctuation would be for written language.

In all of this work, we also consider that word tokenization is given, not that it is a trivial problem in general, but because it is arbitrary and mention detection evaluation depends on it. Therefore, having our system perform word tokenization on its own could potentially add noise to the evaluation that would only be due to having different conventions for word boundaries.

6.1.1 Span features

At the span representation level, there are few useful features of this type: Clark and Manning (2016b) for instance uses the length of the mention in number of words, its absolute position in the document (in number of mentions), inclusion in another mention and its type (pronoun, noun, proper noun or list) and a general document genre embedding. Of these, those that depend on document-level context or pre-existing mention detection and classification are not relevant for span representation, since we obviously do not have access to detected mentions at this stage. Genre embeddings also seem relevant in corpora made of heterogeneous collections, but in the case of ANCOR, given the size distribution of the subcorpora and their relatively close language varieties, (see section 3.2.2), it would be hard to learn and not necessarily very useful. This only leaves the mentions' length, which could theoretically be learnt by LSTM_w (Weiss et al. 2018), but providing it directly to the model is easy enough and relieves LSTM_w of that burden. Raghunathan et al. (2010) also mention a character-based acronym-reconstruction feature, which is certainly useful in acronym-rich corpora, but of little relevance for ANCOR.

As in Clark and Manning (2016b) and K. Lee et al. (2017), we bin span length in the following buckets[2]: $[0, 1[$, $[2, 3[$, $[3, 4[$, $[5, 7[$, $[8, 15[$, $[16, 31[$, $[32, 63[$ and $[64, +\infty[$, in order to densify the training signal somewhat, since there should be few differences in practice between a span of length 9 and one of length 10 for instance. The bin index is then embedded via a lookup in a table of end-to-end trained embeddings. The same procedure is used for all the integer features in our system and could also be a factor in the relevance of this feature: even though LSTM_w can in principle recover the length of a span, deriving a useful dense representation without further supervision is less likely.

[2]Introduced by Clark and Manning (2016b) with no motivation for their boundaries, which are clearly $\lfloor \log_2(\ell) \rfloor$

Table 6.1: Influence of the length feature on mention detection (ANCOR dev)

Setting	P	R	F
Baseline	86.89	88.15	87.52
-length	87.28	84.63	85.94

Table 6.2: Influence of the length feature on coreference resolution (ANCOR dev)

	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
System	P	R	F	P	R	F	P	R	F	Avg.	P	R	F
Baseline	73.60	53.54	61.74	89.40	74.67	81.32	76.12	88.07	81.59	74.88	79.61	64.54	69.46
-length	72.88	50.85	59.52	88.90	73.82	80.60	75.05	88.12	80.98	73.70	76.80	63.91	68.08

Table 6.1 compares the performance for mention detection with and without using the span length feature and table 6.2 those for coreference resolution. These results suggest that this feature is indeed a valuable addition to span representation, despite its seemingly redundant nature.

6.1.2 Pair features

As we saw in section 4.1.2, mention pairs features usually belong to one of the following categories: similarity and agreement features, distance features and compatibility features. Among these, most distance features are raw, since they depend on either basic segmentation or mention detection (which is available at this stage) ; similarity features between raw mention-level features are also obviously raw, but most compatibility features require information from further processing.

In our baseline system, we use the following features:

- Distance in number of words
- Distance in number of utterances
- Distance in number of mentions
- Speaker agreement
- Inclusion of the mention in the antecedent candidate (equivalent to Raghunathan et al. (2010)'s 'i-within-i' feature)

Since we only represent spans in their immediate context, these features are the only access that the antecedent scorer has to the document (and therefore discourse) structure, which is critical to most reference phenomena (see section 2.1.2). For instance, without access to at

with a special case for 3.

least one distance feature, pronouns would have to be attached by choosing randomly between the compatible antecedent candidates. Similarly, without a speaker agreement feature, there would be no way for our system to determine if two self-reference from a speaker can be coreferent. For these reasons, experimenting without these features is unnecessary, since there is no possibility of it resulting in anything but severe performance degradations.

A raw pair feature that is commonly used in existing systems and that is not strictly necessary for our system is string matching. String matching has been a staple of machine-learning systems for coreference resolution since Soon et al. (2001) and until very recently (Clark and Manning 2016b; H. Lee et al. 2017) it was included in almost all of them. Moreover, it is very often one of the most informative features: Soon et al. (2001) reports that only using string matching is already enough to obtain significantly better performances than the previous best machine learning system (RESOLVE, McCarthy and Lehnert 1995) and its ablation in Clark and Manning (2016b) results in the second-worse performance loss (behind length features).

It is quite clear why this information is so valuable: for noun phrases and close pronouns, it is very unlikely that two matching mentions are not coreferent, therefore, trusting this feature —with some restrictions on distance for pronouns— should almost never result in a mistake. On the other hand, even for systems that are able to use word-level representations, learning to match mention representations when they are matching *but do not appear in the same context* is highly non-trivial. For instance, in our implementation (see section 5.2.2), the only way to access this feature would be via the soft-head representation, and it would require attention weights to be independent of the recurrent representations of words. Predicting this feature is clearly not a reasonable investment should we force the network to learn it and it is very unlikely that it would be learned otherwise. Conversely, including this feature explicitly is not very expensive, as it is easily computed.

We experiment with a relaxed version of string matching. Instead of strict string matching as a boolean feature, we use two similarity measures, considering spans as bag of words: the Jaccard similarity coefficient J (Jaccard 1912) and the Szymkiewicz-Simpson overlap coefficient O (Simpson 1947; Szymkiewicz 1934):

$$\begin{aligned} J(A, B) &= \frac{|A \cap B|}{|A \cup B|} \\ O(A, B) &= \frac{|A \cap B|}{\min(|A|, |B|)} \end{aligned} \qquad [6.1]$$

These allow us to represent exact string matching $J(A, B) = 1$, strict inclusion $O(A, B) = 1$ but also relaxations of both cases, in order to take into account small amounts of modifications —for instance the addition of a modifier, changing an article from definite to indefinite or the inclusion of disfluencies—. This also allows us to delegate the task of choosing a sensibility threshold to the system instead of allowing only strict matches.

Table 6.3 shows the results of adding string matching features to our baseline system for coreference resolutions. The gains are obvious for all metrics and are easily interpretable as improvements in correctly predicting coreference, which results in recall gains for MUC, B^3

Table 6.3: Influence of string matching features on coreference resolution (ANCOR dev)

	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
Baseline	73.60	53.54	61.74	89.40	74.67	81.32	76.12	88.07	81.59	74.88	79.61	64.54	69.46
+string	77.76	60.18	67.59	89.66	76.63	82.58	79.76	89.69	84.38	78.18	79.92	66.16	70.78

and BLANC (morally recall gains in coreference identification) and precision gains for $CEAF_e$, indicating less oversplitting of coreference chains.

6.2 Semi-supervised knowledge

We call *semi-supervised knowledge* any information that can be obtained by processing unannotated linguistic data, usually in the target language or sometimes in closely related ones (e.g. written French wrt. spoken French). The term *semi-supervised* (Collobert and Weston (2008) among others) denotes the fact that this knowledge is often obtained from supervised learning methods, hence not strictly *unsupervised* but that supervision does not come from any extra analysis of the data. This kind of information is more easily available than knowledge that requires the development of specific tools, especially since the development of methods to extract it from web content (Grave et al. 2018; Mikolov et al. 2018), which provides very large scale repositories of such data — if not necessarily in clean or reliable forms.

Though the idea of using knowledge extracted from raw data in downstream task is older — it is found for instance in the concept of Brown Clusters (P. F. Brown et al. 1992) —, it is mostly with its integration in neural architectures in the form of pretrained word embeddings, championed by Collobert and Weston (2008), Maas et al. (2011), Mikolov (2012) and Turian et al. (2010) and popularized to omnipresence by Mikolov et al. (2013) that its relevance became evident. It is in this form that we include semi-supervised knowledge in our architecture, both as of *static* pretrained word embeddings (section 6.2.1) and in the form of the more recent *contextual* word embeddings (section 6.2.2). A form of unsupervised knowledge that we did not explore is the unsupervised morphological semantic segmentation proposed by Grönroos et al. (2014), to which the subword segmentation used in BERT (Devlin et al. 2018, see section 6.2.2) is analogue.

6.2.1 Static word embeddings

The hypothesis underlying most word embeddings pretraining is the *distributional semantics hypothesis* that 'words which are similar in meaning occur in similar contexts' (Rubenstein and Goodenough 1965), implying that semantic information about a word can be learned by observing its contexts of occurrence in a large enough corpus. In practice, the most successful implementations rely on some form of language modelling task, where a neural network is trained to predict the likelihood of occurrence of a word given a context (or more rarely the

opposite as with the continuous skip-gram model of Mikolov et al. (2013)), the word embeddings are then extracted from the learned internal representation of words in the network.

While there are many variations on this simple idea, the best known are the word2vec models of Mikolov et al. (2013), the GloVe model of Pennington et al. (2014) —which combine a language modelling objective with a variant of earlier formulations based on the factorization of a co-occurrence matrix— and FastText (Bojanowski et al. 2017; Mikolov et al. 2018) —which applies Mikolov et al. (2013)'s skip-gram model to subword units. In practice, the respective merits of these models are hard to assess: while they are usually evaluated using word analogy or document classification tasks, few works have explored their influence on more involved downstream tasks or cross-domain usage. Furthermore, the availability of pretrained models for languages other than English is not very consistent and the models that do exist often suffer from a lack of extensive documentation of the data and configuration used, which can be an issue, considering the influence of the hyperparameter 'tricks' described by Mikolov et al. (2018).

In consequence, we did not experiment with alternative models and limited ourselves to FastText. We use it with the pretrained embeddings released by Grave et al. (2018) for our baseline as described in section 5.4.1. We do not use the sub-word representations provided by FastText, since they would make the comparison between the two settings less clear. In both settings, the word embeddings are fine-tuned during the training procedure, and the span representations learned for mention detection are carried over for coreference resolution in the *sequential* mode described in section 5.4.3.

Table 6.4 includes a comparison of the mention detection performances of our baseline and the same model, trained in the same setting, but without pretrained word embeddings ; table 6.5 includes the same comparison for coreference resolution. In both cases, the difference is not significant, though mention detection seems somewhat harder without pretrained embeddings. A tentative interpretation for these results could be that while the FastText embeddings help at the beginning of training, the genre difference between the data they were trained for (web text) and the spoken French of ANCOR is sufficient to allow randomly-initialized embeddings to reach similar performances using only the signal in ANCOR. This hypothesis is also consistent with the much slower convergence of the models without pretraining, which took in average 5 more epochs to overfit on the development set for mention detection and 10 more epochs for coreference resolution, an increase of about 50 % in both cases.

6.2.2 Contextual word embeddings

Static word embeddings are usually represented by a mapping from a word *form* to a dense vector, which implies that homographs are represented by the same vector while alternative orthographies of the same word are not. This is symptomatic of larger issue: although they rely on the assumed dependency between the meaning of a word and its context, static word embeddings are —by definition— *context-agnostic* and a static word embedding is in effect a mixture of the representations of all the meanings of the corresponding word form. This goes beyond simple homography, since it also hampers the capacity of a system to take into account e.g. figurative meanings of words.

While some earlier[3] works such as Choi et al. (2017) and Frederick Liu et al. (2018) report efforts to address this issue at a downstream level, the first successful implementation of general-purpose[4] context-dependent word embeddings is ELMo due to Peters et al. (2018). In spirit, it builds upon the same distributional semantics hypothesis as static word embeddings, and it is also implemented via a language modelling task, but instead of pooling word forms representations over all their occurrences, it instead proposes to keep a context-specific representation obtained by computing a weighted average of the internal states of stacked LSTM layers. These contextual representations give a clear advantage for many downstream tasks, as seen in the original article and more specifically in K. Lee et al. (2018), where merely replacing static word embeddings by ELMo representations improves the result of K. Lee et al. (2017)'s baseline by more than 3 F_1 percentage points for all the CoNLL metrics.

The cost for these impressive performance gains is a far worse computational cost, since it replaces a simple table lookup for static word embeddings by a sentence- or document-wise processing by a mutilayer recurrent neural network. However, fine-tuning ELMo for downstream tasks only requires learning a very small number of parameters compared to fine-tuning static word embeddings, since the parameters of the recurrent layers themselves can be frozen and only the averaging weights have to be tuned. As a result, while using ELMo representations tends to be slower at inference time, in our experiments, it was not noticeably slower than the alternatives.

Despite its success and fast adoption by the natural language processing community, there have not been a lot of works using ELMo in coreference resolution, as it has soon been supplanted by BERT (Devlin et al. 2018) and its successors. In particular, alternative pretrained ELMo models are uncommon, especially for languages other than English. In our case, we experiment with an ELMo model trained on the French part of OSCAR (Ortiz Suárez et al. 2019) for Ortiz Suárez et al. (2020).[5]

BERT (Devlin et al. 2018) is a similar proposal, that replaces the bidirectional language model objective of ELMo by a masked language model, and its stacked LSTM architecture by a deep stack of Transformer blocks (Vaswani et al. 2017). Its main advantage over ELMo resides in this substitution, which allows a much more efficient training, since the feedforward architecture of Transformers is easily parallelizable, especially using dedicated hardware (Jouppi et al. 2017). This in turn allows for much larger modelsitems, trained on much larger datasets and extensive fine-tuning on multiple downstream tasks that would be impractical with ELMo. Consequently, the performances of all the currently published BERT-based models outperform the ELMo- and static embeddings-based alternatives, often by a fair margin, and coreference resolution is not different in that respect (Joshi et al. 2019b).

Due to the prohibitive cost of pre-training BERT, models for languages other than English are still not very common, but two pretrained models exist for French: CamemBERT (L. Martin et al. 2020) and FlauBERT (H. Le et al. 2020). We experiment with CamemBERT — which is also trained

[3] Or, to be precise, *published simultaneously* in the case of Frederick Liu et al. (2018).

[4] Dai and Quoc V Le (2015) proposed a very similar idea, but with task-dependent embeddings.

[5] Many thanks to Pedro Ortiz Suárez for giving us access to this model ahead of time.

on the French part of OSCAR and is therefore more comparable with our ELMo representations — and the mutlilingual version of BERT provided by Devlin et al. (2018) (mBERT). Following Joshi et al. (2019b), we also made a small change in our architecture to accommodate for BERT subword tokenization: instead of using the bounding *words* in our spans representations, we use the bounding word *pieces*.

Table 6.4: Influence of word embeddings on mention detection (ANCOR dev)

Setting	**P**	**R**	**F**
No pretraining	84.68	88.89	86.73
Baseline (FastText)	86.89	88.15	87.52
ELMo	89.67	91.96	90.80
mBERT	84.06	90.50	87.16
CamemBERT	88.47	93.24	90.79

Table 6.5: Influence of word embeddings on coreference resolution (ANCOR dev)

	MUC			**B^3**			**$CEAF_e$**			**CoNLL**	**BLANC**		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
No pretraining	73.49	54.35	62.15	88.64	75.15	81.24	76.11	87.45	81.29	74.89	76.80	65.72	69.50
Baseline (FastText)	73.60	53.54	61.74	89.40	74.67	81.32	76.12	88.07	81.59	74.88	79.61	64.54	69.46
ELMo	76.33	60.21	67.06	88.67	77.30	82.53	79.24	88.53	83.56	77.72	78.90	67.84	71.74
mBERT	73.93	56.76	64.02	88.90	75.65	81.67	77.68	87.87	82.40	76.03	79.88	65.66	70.42
CamemBERT	76.20	60.61	67.32	88.52	77.40	82.53	79.40	88.47	83.63	77.83	79.27	67.94	71.96
ELMo+string	79.93	66.58	72.50	89.92	79.32	84.24	82.69	90.13	86.21	80.98	82.67	69.39	74.15

We report in table 6.4 a comparison between all the word embedding techniques that we have described in this section for mention detection, and in table 6.5 for coreference resolution. In all cases, contextual embeddings outperform static embeddings and non-pretrained embeddings, by a fair margin for the monolingual models and by a smaller one for mBERT (which actually perform worse, if not significantly so on mention detection).

Surprisingly, CamemBERT does not bring any significant improvement over ELMo, contrarily to the situation for English reported by Joshi et al. (2019b). An explanation for this might reside in their respective input representations: ELMo uses a character-level convolutional neural layer, which is robust to both out-of-vocabulary words and irregular spelling, while CamemBERT relies on subwords obtained by optimizing a byte pair encoding vocabulary (Sennrich et al. 2016) for its source corpus, in that case a web-based one. As a result, the CamemBERT tokenizing model is less robust to the specific spellings used in speech transcriptions, in particular for incomplete words: it tokenizes 'Je suis con- euh concentré' as 'Je', 'suis', 'con', '-,' 'eu', 'h,' 'con', 'cent', 'ré' for instance, which introduces a false homography between the 'eu' part of the interjection 'euh' and the 'eu' past participle form of 'avoir' (despite having had exposure to 'euh' in its training data), while ELMo correctly treats 'euh' as its own token.

Finally, we also include the result of augmenting our baseline with both ELMo representations and the string matching features described in section 6.1.2, which shows the same gain over simply using ELMo as using the string features does over the baseline. This suggests that the gain brought by ELMo representations is not simply due to the differences in their representations of inputs, but comes from additional knowledge learned from their pretraining for language modelling on a large corpus.

6.3 Shallow linguistic knowledge

We define *shallow linguistic knowledge* as knowledge that is ultimately derived from annotated resources and concerns the nature of linguistic material, but not its structure. In particular, we do not include knowledge that concerns segments not present in the raw material (i.e. words and utterances in the case of ANCOR) in this category. The most common types of shallow linguistic knowledge are word-level syntactic and lexical features, e.g. POS annotations, lemmas and morphological features, but semantic information such as sentiment polarity and sentence-level dialogue act features are also part of this category.

In our case, we experiment with two instances of shallow linguistic knowledge: POS tags and lemmas. These are two of the most common and simple shallow features found commonly in natural language processing systems, POS tags in particular are commonly used in existing coreference resolution systems, often as part of a syntactic parsing pipeline. POs tagging is also a relatively well understood task, and current automatic systems commonly reach human-level performances on this task in in-domain evaluations for well-resourced languages (Bohnet et al. 2018; Heinzerling and Strube 2019) although the performances for less regular genres such as social media texts tends to be somewhat worse (Godin 2019).

Lemmas are less universally used than POS features, since for morphologically-poor languages (e.g. English), they are not much more informative than a combination of POS and word embeddings, which are easier to obtain. Proper lemmatization of morphologically-rich languages and non-standard varieties, however, is still very much an open problem (Manjavacas et al. 2019) and is still relevant as an extreme version of word form normalization, which can have a non-negligible influence on downstream tasks (Straka and Straková 2019).

Specifically, in our experiments, we use the POS tags and lemmas provided by spaCy (Honnibal and Montani 2019) —precisely with the `fr_core_news_sm` version 2.2.5 model, trained on the Sequoia (Candito and Seddah 2012) and WikiNER (Nothman et al. 2013) corpora. In terms of performances, this model is somewhat behind the current state-of-the for French, however it has the advantage of being very fast and easy to integrate in a preprocessing pipeline. In any case, since there are currently no standard evaluation dataset for POS-tagging and lemmatization for spoken French, evaluating the relative performance of systems designed for written French on these auxiliary tasks would be out-of-scope for this work. A cursory exploration of the outputs of this model on ANCOR suggest that the POS and lemma, while obviously not perfectly accurate, were generally correct for parts with a low amount of disfluencies and low interactivities but more brittle in the most interactive parts. Conversely, although this model also provide additional morphological features (in the Universal Dependency (Nivre et al. 2016) formalism),

we did not experiment with these, as they were usually less reliable, harder to integrate in our architecture (since they are in the form of variable-length multi-valued features) and since the information they provide seemed too sparse to be useful.

In both cases, we augment our baseline architecture with these features by simply concatenating their embeddings with the word representations $e(w)$ given by our word-encoding layer (see section 5.2.2). These embeddings are initialized randomly —in particular, lemma embeddings do not share their weights with the word embeddings— and are of dimension 50 for the POS and 150 for the lemmas.

Table 6.6: Influence of shallow linguistic knowledge on mention detection (ANCOR dev)

Setting	**P**	**R**	**F**
Baseline	86.89	88.15	87.52
+POS	87.24	89.62	88.41
+lemmas	86.65	89.59	88.09X

Table 6.7: Influence of shallow linguistic knowledge on coreference resolution (ANCOR dev)

	MUC			**B^3**			**$CEAF_e$**			**CoNLL**	**BLANC**		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
Baseline	73.60	53.54	61.74	89.40	74.67	81.32	76.12	88.07	81.59	74.88	79.61	64.54	69.46
+POS	70.95	50.24	58.61	89.22	73.42	80.48	74.83	87.55	80.62	73.24	79.05	63.18	68.11
+lemmas	74.09	50.40	59.71	90.14	73.73	81.06	74.61	88.50	80.90	73.89	78.69	63.97	68.66

We report in table 6.6 the comparisos between our baseline and its augmentation with shallow linguistic knowledge for mention detection and in table 6.7 for coreference resolution. For both tasks, the changes brought by these augmentations are not significant, and their interpretation can therefore only be speculative. However, the somewhat better results for mention detection with POS could be an indication that the span representation module can learn to use this information for mention detection, and conversely that learning to rely on it too much can reduce its effectiveness for coreference resolution, where POS are less relevant.

6.4 Structural linguistic knowledge

In opposition to *shallow* linguistic knowledge (section 6.3), we call *structural linguistic knowledge* any information that both ultimately derives from annotated resources and provide structural information on linguistic material. Structural information can be non-recursive, for segmentation information beyond those provided by the base material —as with chunking and named entity boundaries— or recursive, the most prominent form of which is syntactic parsing.

The main advantage of structural knowledge over the other types of information we experiment with in this chapter is that in many cases, the additional segmentation that it provided is correlated with mention boundaries: named entities are mentions, and in ANCOR non-nested

mentions are nominal chunks and all mentions are either pronouns or noun phrases. Consequently, successfully integrating this kind of knowledge as part of our mention detection architecture should relieve it of having to deal with the simple cases where these features are sufficient and let it concentrate instead on more complex cases.

However, given that the tools that we use to obtain this knowledge were not designed for spoken French, it is to be expected that the segmentations they provide will not be completely accurate, or even sufficiently accurate to use them as a resource for rule-based processing. Instead, we use it merely as an additional source of information, and include not only the exact boundaries that they provide but also features based on partial matching with span boundaries.

In this section, we report our experiments on integrating features obtained from a chunker, a named entity detector and a syntactic parser. In all cases, we integrate these features in our system in two ways: as span level features and as a guide for the undersampling performed on the training set (see section 5.3.4).

6.4.1 Non-recursive structures

We experiment with two types of information on non-recursive structures: chunks and named entities.

Chunking, introduced by Abney (1992) as an intermediary step in the development of a syntactic parser consists in segmenting a sentence in non-overlapping, non-nested syntactic units, such as those in the following bracketing '[I begin] [with an intuition]: [when I read] [a sentence], [I read it][a chunk] [at a time]' (Abney 1992). While Abney (1992) does not describe it as an end it itself, chunking has subsequently become a very popular lightweight substitute for full syntactic parsing. Indeed, the non-recursive nature of chunking allows to cast it a sequence labelling task, for which efficient algorithms are known, in both rule-based and machine learning configurations (Sha and Pereira 2003).

For coreference as annotated in ANCOR, the chunks of interest are *noun* chunks, since their boundaries coincide with those of non-nested mentions. Works such as Broscheit et al. (2010) and Soraluze et al. (2012), operating in context where syntactic parsing is unavailable or unreliable have successfully used chunking in combination with handcrafted rules (in order to recover the nested structure of mentions) for mention detection.

We experiment with integrating the result of chunking in our system in two ways. First, every span receives an additional feature depending on its overlapping with one of the nominal chunks detected for its sentence, this feature can take one of the four following values:

`exact` if the span boundaries correspond exactly to those of a nominal chunk

`included` if the span is a proper subsequence of a nominal chunk

`outside` if the span does not intersect with any nominal chunk

`incompatible` if the span intersects with a nominal chunk, but neither is a subsequence of the other

Second, while undersampling for non-mention spans for the training set of our system, we prevent noun chunks from being discarded. The rationale behind this choice is that detected non-mention nominal chunks should be harder to distinguish from mention spans than any arbitrary span, since these spans should be more syntactically similar. Our hope is that leaning this harder task might make our mention detection module perform better overall.

As for named entities, the theoretical (section 2.3.1) and practical (section 4.1.1) links between named entity recognition and coreference resolution are clear. Since named entities are a specific class of mention with a specific mode of reference, having access to their boundaries and their types should in principle be a valuable asset for mention detection, and possibly for coreference resolution. Accordingly, we enhance our spans representations with a categorical feature whose value for named entity spans is the type of the corresponding entity and `None` for the other spans. Contrarily to chunks, in order to maximize the precision of this feature, we only consider exactly matching spans and do not report partial overlaps, but we do prevent erroneously detected named entity spans from being discarded during undersampling.

As in our experiments with shallow linguistic knowledge (section 6.3), we use spaCy for named entity noun chunk detection, in the same configuration as previously. Other chunkers exist for French, such as SEM (Dupont 2017; Tellier et al. 2012), but none for spoken French[6] —hence our choice of a somewhat worse preprocessing in terms of accuracy but one that more easily integrable in our pipeline. The situation for named entity recognition is somewhat better, as several named entity recognition systems were developed for broadcast spoken French on the occasion of the ESTER 2 (Galliano et al. 2009) and ETAPE (Gravier et al. 2012) and in independent works such as Hatmi (2014) and Zidouni et al. (2010), but none of these systems was both publicly available and easily integrable in our pipeline.

Finally, similarly to word-level features both of these features are represented in our system by concatenating a feature embedding (of dimension 16) to the spans representations. These embeddings are also randomly initialized and learned during the training phase.

6.4.2 Syntactic parsing

Syntax, in one form of another is one of the oldest concerns in linguistics[7] and the analytic tools it provides are the foundation of many linguistic theories, including theories of anaphora, such as the syntactic formulation of the government and binding theory of pronominal anaphora (Chomsky 1981). It is thus only natural that syntactic parsing was a prerequisite to most of the coreference resolution systems until very recently (see chapter 6). Indeed, in most models of syntax and reference, mentions coincide with syntactic objects —usually constituents or equivalents and their roles in the syntactic structure of sentences has a clear influence on the reference identification process. Therefore, for mention detection, having access to constituents allows for considerable simplifications, with relatively simple heuristics (see section 4.3) providing strong baseline strong baselines. As for coreference resolution, beyond the integration of

[6] Although the results of Tellier et al. (2013, 2014) suggest that a bootstrapping procedure could yield one with limited efforts, designing such a system is beyond the scope of this work.

[7] With roots as old as Pāṇini (–0349).

syntactic constraints and preferences, many systems also use rules or features that relies on the availability of mention heads (section 4.1.1), implicitly relying on the strong correlation between the semantic content of the head of a constituent and that of the constituent itself.

In our case, we use the dependency syntax analyses (in the Universal Dependencies formalism) included in ANCOR-AS (Grobol et al. 2018b), obtained from the then-state-of-the-art parser for written French Dyalog-SRNN (Villemonte De La Clergerie et al. 2017). As equivalents to noun phrases, we extract noun-headed subtrees from these analyses, using simple heuristics to account for specific idiosyncrasies of the annotation scheme —most notably the choice of using the subject as the root of copular constructions.

As with chunks, in order to account for errors in these analyses, we also introduce a relaxation in the matching between spans and subtrees: we define the distance between a span $w_i, ..., w_j$ and a subtree of maximal projection $w_k, ..., w_\ell$ as $|i-k| + |j-\ell|$ and associate every span with the closest noun-headed subtree. Using this matching, we then enhance our spans representations with the following features, also added as embeddings to span representations:

- Head word form (dimension 300)
- Head POS (dimension 50)
- Head dependency type (dimension 50)
- Head's governor word form (dimension 300)
- Head's governor POS (dimension 50)
- Distance to the corresponding span (dimension 20)

In addition to our usual setting for mention detection where all spans up to a certain length are considered as mention candidates, we also experiment with an alternative scheme where mention candidates are restricted to the span within a certain distance of their associated span. Unlike the restriction to undersampling that we use for chunks and named entities, this scheme is not only applied to the training set but also to the development and test set. This scheme thus serves as a filter that reduces the class imbalance for mention detection, by removing many spurious mention candidates —which should make the mention detection task easier while still leaving the possibility of recovering from minor parsing errors. However, for spans with severe parsing errors, mention spans could still be lost, depending on the maximal allowed distance between spans and subtrees. In our experiments, we chose to limit this distance to 3, which covers about 96.5 % of the mention spans. This alternative scheme is noted by 'tree' in table 6.8 and table 6.9, while our baseline scheme is noted by 'span'.

6.4.3 Results

We report in table 6.8 the comparison of the structural linguistic features described in this section. Clearly, in our experiments, adding information on noun chunk and on named entities based on external preprocessing tools does not have a significant impact on mention detection. These results suggest that either our system is able to learn implicitly the knowledge that the

Table 6.8: Influence of structural linguistic knowledge on mention detection (ANCOR dev)

Setting	**P**	**R**	**F**
Baseline	86.89	88.15	87.52
+chunk	87.20	88.03	87.61
+NER	86.28	88.26	87.26
+parsing (tree)	91.32	83.76	87.38
+parsing (span)	86.59	90.49	88.21

tools provide using only the information latent in our training data, or alternatively, that these tools are too unreliable in this cross-domain setting to allow our system to learn from them.

The case of syntactic parsing is more interesting: in the tree-based setting, although the final F_1 is not significantly different from the baseline's, it comes from different scores, with both a significantly better precision and a significantly worse recall. This is consistent with our hypothesis that relaxed subtrees extracted from a syntactic analysis can be an interesting alternative to our baseline of using every spans, as it tends to give less imbalance between mentions and non-mentions. Conversely, it also confirms that the accuracy of the syntactic analysis used for this is critical, since it makes some mention spans irrecoverable, which can significantly increase the false negative rate.

The results in the span-based setting are better, and significantly so for recall, which suggests that there is indeed some knowledge captured by syntactic parsing that our system does not learn by itself and confirms our intuition that there should be a way to take advantage of external structured linguistic knowledge. The fact that these better results were obtained using knowledge from a state-of-the-art system rather than the 'good enough' tools used for the other features also tends to confirm that the cross-domain relevance of preprocessing tools are indeed correlated to their relative in-domain performances.

Table 6.9: Influence of structural linguistic knowledge on coreference resolution (ANCOR dev)

	MUC			**B^3**			**$CEAF_e$**			**CoNLL**	**BLANC**		
System	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg.**	**P**	**R**	**F**
Baseline	73.60	53.54	61.74	89.40	74.67	81.32	76.12	88.07	81.59	74.88	79.61	64.54	69.46
+chunks	72.46	51.94	60.19	88.26	74.29	80.61	75.67	87.97	81.28	74.03	75.92	64.66	68.43
+NER	74.19	47.60	57.69	90.93	72.37	80.52	73.29	88.70	80.19	72.80	80.44	62.16	67.36
+parsing	73.25	51.32	60.07	89.84	73.70	80.90	75.10	88.04	80.98	73.98	80.17	63.61	68.64

Table 6.9 gives a comparison of the influence of structural linguistic features on coreference resolution. As for mention detection, the differences with our baseline are not significant, suggesting that these features are not informative or reliable enough to be of use for our system.

This is not very surprising for chunking: since this evaluation is on gold mentions, the only information that noun chunk boundaries could bring to this task would be on cases where the chunker was consistently failing to detect some categories of chunks, which does not seem like a reasonable hypothesis. As for named entities their boundaries are not more relevant than those of noun chunks, and although their categorization might have been useful in other domains where named entities appear in coreference chains with other noun phrases, named entity mentions are probably too few —both in occurrences and in unique entities— in ANCOR for it to make a significant difference.

The result for syntactic parsing are more disappointing. Given the importance of head-matching features in other coreference resolution systems —including systems based on neural networks such as Clark and Manning (2016a) and Wiseman et al. (2015)— and the success of syntax-based heuristics in earlier works, one could have hoped to see more improvements by integrating explicit syntactic features in our system. However, the correlation noted by K. Lee et al. (2017) between the representations obtained via a soft-head attention mechanism trained for coreference resolution and the syntactic heads of mentions could suggest that this kind of information is recoverable without relying on a syntactic parser. In fact, it might actually be that the domination of the head in the semantic content of span and its importance in coreference resolution makes this knowledge inferrable using only coreference information.

6.5 Conclusion

In this chapter, we studied the impact of adding various forms of knowledge to the baseline system presented in chapter 5 on its performances. We find that in general, knowledge obtained from preprocessing tools —on which all coreference resolution systems were based until very recently— does not significantly improve nor degrades the performances of our system. Conversely, we find that knowledge derived from semi-supervised representation learning can greatly improve these performances, as does the addition of very simple string-matching features that do not require external resources.

However, these results come with some caveats: few preprocessing tools for spoken French are currently available —and almost none for the *spontaneous* genre used in ANCOR— which led us to use tools designed for *written* French instead. In consequence, although we had no way to quantify it, a degradation of their performances was to be expected, especially since the specific tools we chose were not specifically designed for cross-domain robustness either. From this, we can conclude that in order for explicit linguistic knowledge to be relevant in this setting, it has to be more accurate and that taking advantage of noisy inputs in neural architectures is not a trivial task. That said, the results reported in section 6.4.2 by using a syntactic parser to improve mention detection gives us hope that using more accurate preprocessing tools would indeed lead to real improvements and in the meantime, that they can be useful in specific contexts (in that case, for downstream applications where the precision of mention detection is more important than its recall).

The good results obtained by using semi-supervised knowledge is not very surprising, since they mostly confirm a trend observed in almost all areas of natural language processing —in

particular for the contextual input embeddings—, but it is interesting to see that contrarily to knowledge extracted from preprocessing tools, the cross-domain setting is not an obstacle here. This might be due to the fact that these representations were specifically designed and trained to be used in neural architectures, with few constraints from linguistic theories, while the information provided by preprocessing tools is necessarily geared toward specific models of language that might not apply to other domains. Another explanation is that the usefulness of these representations comes from the possibility of leveraging far more linguistic data than is available for tools that need access to annotated corpora. Nevertheless, let us note that these gains in performance are not free or even cheap: these representations learning techniques *require* consequent amounts of data and computational power in order to reach these levels of accuracy, which would be problematic for uses in lesser-resourced languages and for groups with limited access to the necessary infrastructures.

Chapter 7

Conclusion and perspectives

In this work, we have dealt with coreference resolution applied to the case of spontaneous spoken French. Our main contribution to this question is DeCOFre, the first end-to-end coreference resolution for French, which also happens to be the first end-to-end coreference resolution system for the spontaneous spoken genre in any language. To overcome the scarcity of resources for spoken French, this system does not depend on external knowledge, and still obtains reasonably good results, quantitatively speaking. We also provide an assessment of the interest of introducing resources developed for written French in that system, and report that while resources built on unannotated corpora can improve on our baseline system, more traditional preprocessing tools have a very limited impact, stressing the importance of developing natural language processing systems that are robust in more genres and domains. Finally, we also propose a new standard way of representing annotated resources that is suitable for coreference corpora, a representation used in the recently released DEMOCRAT corpus (Landragin 2016) and an improved version of ANCOR (Muzerelle et al. 2013a, 2014), the resources on which this work is based.

In the short to immediate term, the clear priority for extending this work should be to reproduce our experiments on the DEMOCRAT corpus, which would make the respective influence of the genre and of the system architecture in DeCOFre performances clearer. To this end, continuations of our standardization effort for coreference corpora in French are underway, and a resource unifying DEMOCRAT and ANCOR will be available soon.

Developing tools and resources specific to spoken French would also make our results from chapter 6 more definitive. It would also be interesting to extend these experiments to other forms of external knowledge, and in particular knowledge based on the *actual* source material (i.e. the audio signal) rather than transcriptions. From a pure quantitative point of view, there is probably a large margin of improvement to be found by exploring the hyperparameter space of our system, as well as by experimenting with other coreference resolution algorithms. That said, if results from other fields in natural language processing are any indication, the easiest way to the raw quantitative performances of our system is probably to improve its input representations, for instance by pretraining contextual word embeddings on spoken French and fine-tuning them on tasks related to coreference resolution.

Although more research is needed in that direction, it seems clear from our experiments that including high-level linguistic knowledge in a neural coreference resolution system is not trivial, especially if this knowledge is unreliable: so far, the capacity of neural networks to make sense of information hidden in noisy inputs does not seem to be enough. However, despite the disappointing results of our experiments with joint learning, other works (such as Fei Liu et al. (2019), Sanh et al. (2018) and Swayamdipta et al. (2018)) suggest that joint learning of tasks of different level could help in that matter. The case of the Referential Reader (Fei Liu et al. 2019), of generative methods in neural parsing (Crabbé et al. 2019; Dyer et al. 2016) and of the various ways to learn representations from language modelling also suggest that semi-supervised tasks are a sensible way to obtain linguistic knowledge in forms that are well suited for neural architectures without requiring expensive resources. In the coming years, this could be a way to help fighting Durrett and Klein (2013)'s 'uphill battles'.

Finally, the successes of the transition-based mention detection system of B. Wang et al. (2018) and of the document-order processing of the Referential Reader (Fei Liu et al. 2019) suggest that the hypothetical document-reading device of Karttunen (1976) might indeed become a practical idea in the near future. More concretely: a multi-task system that would process a document in reading order, with explicit entity modelling and joint parsing and mention detection could be a way to improve on the current models in terms of sheer performance, interpretability and consistency with cognitive models of coreference.

Finally, drawing inspiration from recent advances in syntactic parsing (Meechan-Maddon and Nivre 2019; Smith et al. 2018), improvements in coreference resolution for lesser-resourced languages and genres could come from jointly learning on multilingual and multigenre datasets. Generally speaking, coreference resolution as a field would greatly benefit of moving toward unified frameworks and representations akin to those proposed by the Universal Dependencies project. A move in this direction was made on the occasion of the multilingual coreference detection shared task at SemEval 2010 (Recasens et al. 2010), which could serve as a starting point for a more unified understanding of coreference resolution across languages.

Bibliography

Abney, Steven P. (1992). 'Parsing By Chunks'. In: *Principle-Based Parsing: Computation and Psycholinguistics*. Ed. by Robert C. Berwick, Steven P. Abney and Carol Tenny. Studies in Linguistics and Philosophy. Dordrecht: Springer Netherlands, pp. 257–278. DOI: 10.1007/978-94-011-3474-3_10. URL: https://doi.org/10.1007/978-94-011-3474-3_10 (visited on 17/02/2020) (cit. on p. 105).

Ackerman, Lauren (2019). 'Syntactic and cognitive issues in investigating gendered coreference'. In: *Glossa*. DOI: 10.5334/gjgl.721. URL: https://eprint.ncl.ac.uk/248876 (visited on 23/01/2020) (cit. on p. 56).

Amsili, Pascal and Olga Seminck (2017). 'Schémas Winograd en français: une étude statistique et comportementale'. In: *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2017. Orléans, France, June 2017. URL: https://hal.archives-ouvertes.fr/hal-01628342 (visited on 30/11/2019) (cit. on p. 18).

Antoine, Jean-Yves (2004). 'Résolutions des anaphores pronominales : quelques postulats du TALN mis à l'épreuve du dialogue oral finalisé'. In: *Actes de la 11ème Conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2004. Fez, Morocco: Association pour le Traitement Automatique des Langues. URL: http://www.info.univ-tours.fr/~antoine/articles/2004_TALN.pdf (visited on 11/11/2019) (cit. on pp. 3, 17, 56).

Aone, Chinatsu and Scott William (1995). 'Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies'. In: *33rd Annual Meeting of the Association for Computational Linguistics*. ACL 1995. Cambridge, Massachusetts, USA: Association for Computational Linguistics, June 1995, pp. 122–129. DOI: 10.3115/981658.981675. URL: https://www.aclweb.org/anthology/P95-1017 (visited on 09/02/2020) (cit. on pp. 51, 60).

Asimov, Isaac (1950). *I, Robot*. Gnome (cit. on p. 1).

Ba, Jimmy Lei, Jamie Ryan Kiros and Geoffrey E. Hinton (2016). 'Layer Normalization'. In: Barcelona, España, 8th Dec. 2016. arXiv: 1607.06450. URL: http://arxiv.org/abs/1607.06450 (visited on 03/03/2019) (cit. on pp. 81, 82).

Bagga, Amit and Breck Baldwin (1998a). 'Algorithms for Scoring Coreference Chains'. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Workshop on Linguistics Coreference. Granada, España: European Language Resource Association, 25th May 1998, pp. 563–566. URL: Workshop%20on%20Linguistics%20Coreference (cit. on p. 23).

Bagga, Amit and Breck Baldwin (1998b). 'Entity-Based Cross-Document Coreferencing Using the Vector Space Model'. In: *Proceedings of COLING 1998: The 17th International Conference*

on Computational Linguistics. ACL-COLING 1998. Vol. 1. Montréal, Canada: Association for Computational Linguistics, 10th Aug. 1998. URL: https://www.aclweb.org/anthology/C98-1012 (visited on 30/11/2019) (cit. on p. 16).

Balazs, Jorge and Yutaka Matsuo (2019). 'Gating Mechanisms for Combining Character and Word-level Word Representations: an Empirical Study'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL 2019 Student Research Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 110–124. DOI: 10.18653/v1/N19-3016. URL: https://www.aclweb.org/anthology/N19-3016 (visited on 11/09/2019) (cit. on p. 79).

Baldwin, Breck, Tom Morton, Amit Bagga, Jason Baldridge, Raman Chandraseker, Alexis Dimitriadis, Kieran Snyder and Magdalena Wolska (1998). 'Description of the UPENN CAMP System as Used for Coreference'. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. MUC 1998. URL: https://www.aclweb.org/anthology/M98-1022 (visited on 19/12/2019) (cit. on p. 31).

Ballesteros, Miguel, Chris Dyer and Noah A. Smith (2015). 'Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 349–359. DOI: 10.18653/v1/D15-1041. URL: http://aclweb.org/anthology/D15-1041 (visited on 03/09/2019) (cit. on p. 72).

Bański, Piotr, Bertrand Gaiffe, Patrice Lopez, Simon Meoni, Laurent Romary, Thomas Schmidt, Peter Stadler and Andreas Witt (2016). 'Wake up, standOff!' Sept. 2016. URL: https://hal.inria.fr/hal-01374102 (visited on 03/07/2017) (cit. on p. 44).

Barras, Claude, Edouard Geoffrois, Zhibiao Wu and Mark Liberman (1998). 'Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech'. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. LREC 1998 (Granada, España). European Language Resources, May 1998. URL: http://trans.sourceforge.net/articles/Transcriber-LREC1998.pdf (visited on 06/07/2017) (cit. on pp. 39, 44).

Baude, Olivier and Céline Dugua (2011). '(Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?' In: *Corpus*. Varia 10.10 (1st Nov. 2011), pp. 99–118. URL: http://journals.openedition.org/corpus/2036 (visited on 10/01/2020) (cit. on p. 37).

Belkin, Mikhail, Daniel Hsu, Siyuan Ma and Soumik Mandal (2018). *Reconciling modern machine learning practice and the bias-variance trade-off*. 28th Dec. 2018. arXiv: 1812.11118 [cs, stat]. URL: http://arxiv.org/abs/1812.11118 (visited on 27/09/2019) (cit. on p. 81).

Bender, Emily M. (2019). 'The #BenderRule: On Naming the Languages We Study and Why It Matters'. In: *The Gradient* (15th Sept. 2019). URL: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/ (visited on 20/01/2020) (cit. on pp. 26, 34).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent and Christian Janvin (2003). 'A neural probabilistic language model'. In: *The Journal of Machine Learning Research* 3 (1st Mar. 2003), pp. 1137–1155. URL: http://dl.acm.org/citation.cfm?id=944919.944966 (visited on 27/09/2019) (cit. on pp. 54, 70).

Bengtson, Eric and Dan Roth (2008). 'Understanding the Value of Features for Coreference Resolution'. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language*

Processing. EMNLP 2008 (Honolulu, Hawai'i, USA). Association for Computational Linguistics, Oct. 2008, pp. 294–303. URL: https://www.aclweb.org/anthology/D08-1031 (visited on 29/01/2020) (cit. on pp. 52, 61, 85).

Bergstra, James S., Rémi Bardenet, Yoshua Bengio and Balázs Kégl (2011). 'Algorithms for Hyper-Parameter Optimization'. In: *Advances in Neural Information Processing Systems*. NeurIPS 2011 (Granada, España). Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger. Vol. 24. Curran Associates, Inc., pp. 2546–2554. URL: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf (visited on 06/10/2019) (cit. on p. 93).

Björkelund, Anders and Jonas Kuhn (2014). 'Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL 2014 (Baltimore, Maryland). Vol. 1. Association for Computational Linguistics, pp. 47–57. DOI: 10.3115/v1/P14-1005. URL: http://aclweb.org/anthology/P14-1005 (visited on 03/03/2019) (cit. on pp. 57, 62).

Bohnet, Bernd, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler and Joshua Maynez (2018). 'Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL 2018 (Melbourne, Australia). Vol. 1. Association for Computational Linguistics, July 2018, pp. 2642–2652. DOI: 10.18653/v1/P18-1246. URL: https://www.aclweb.org/anthology/P18-1246 (visited on 17/02/2020) (cit. on p. 103).

Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov (2017). 'Enriching Word Vectors with Subword Information'. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://www.aclweb.org/anthology/Q17-1010 (visited on 16/02/2020) (cit. on p. 100).

Brassier, Maëlle, Alexis Puret, Augustin Voisin-Marras and Loïc Grobol (2018). 'Classification par paires de mention pour la résolution des coréférences en français parlé interactif'. In: *Actes de la Conférence jointe CORIA-TALN-RJC 2018*. CORIA-TALN-RJC 2018. Rennes, France: Association pour le Traitement Automatique des Langues, 14th May 2018. URL: https://hal.inria.fr/hal-01821213/document (visited on 30/01/2019) (cit. on p. 67).

Broscheit, Samuel, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley and Roberto Zanoli (2010). 'BART: A Multilingual Anaphora Resolution System'. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval 2010 (Uppsala, Sverige). Association for Computational Linguistics, pp. 104–107. URL: http://dl.acm.org/citation.cfm?id=1859664.1859685 (visited on 29/03/2018) (cit. on pp. 50, 52, 105).

Brown, Gillian and George Yule (1983). *Discourse Analysis*. Cambridge University Press, 28th July 1983. 210 pp. Google Books: ZUnEAgAAQBAJ (cit. on p. 9).

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai and Robert L. Mercer (1992). 'Class-Based n-gram Models of Natural Language'. In: *Computational Linguistics* 18.4, pp. 467–480. URL: https://www.aclweb.org/anthology/J92-4003 (visited on 16/02/2020) (cit. on p. 99).

Bruneseaux, Florence and Laurent Romary (1997). 'Codage des références et coréférences dans les DHM'. In: *Proceedings of the Joint International Conference of the Association for Computers*

and the Humanities and the Association for Literary & Linguistic Computing. ACH-ALLC '97 (3rd–7th June 1997). Kingston, Ontario, Canada: Association for Computers and the Humanities. URL: https://hal.inria.fr/inria-00433399 (cit. on pp. 36, 43).

Buda, Mateusz, Atsuto Maki and Maciej A. Mazurowski (2018). 'A systematic study of the class imbalance problem in convolutional neural networks'. In: *Neural Networks* 106 (1st Oct. 2018), pp. 249–259. DOI: 10.1016/j.neunet.2018.07.011. URL: http://www.sciencedirect.com/science/article/pii/S0893608018302107 (visited on 21/09/2019) (cit. on p. 83).

Burnard, Lou (2014). *What is the Text Encoding Initiative? : How to add intelligent markup to digital resources*. Encyclopédie numérique. Marseille: OpenEdition Press, 11th Apr. 2014. 114 pp. URL: http://books.openedition.org/oep/426 (visited on 22/01/2020) (cit. on p. 43).

Cai, Jie and Michael Strube (2010). 'End-to-End Coreference Resolution via Hypergraph Partitioning'. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING 2010 (Beijing, China). Aug. 2010, pp. 143–151. URL: https://www.aclweb.org/anthology/C10-1017 (visited on 09/02/2020) (cit. on p. 59).

Calzolari, Nicoletta et al. (2010). 'The LREC Map of Language Resources and Technologies'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. LREC 2010 (Valletta, Malta). European Language Resources Association, May 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/370_Paper.pdf (visited on 14/07/2020) (cit. on p. 35).

Candito, Marie and Djamé Seddah (2012). 'Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical'. In: *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*. TALN 2012 (Grenoble, France). Vol. 2. Association pour le Traitement Automatique des Langues, June 2012. URL: https://hal.inria.fr/hal-00698938 (visited on 17/02/2020) (cit. on p. 103).

Chen, Mia Xu et al. (2018). 'The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL 2018 (Melbourne, Australia). Vol. 1. Association for Computational Linguistics, pp. 76–86. URL: http://aclweb.org/anthology/P18-1008 (visited on 16/02/2019) (cit. on p. 83).

Chetlur, Sharan, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro and Evan Shelhamer (2014). *cuDNN: Efficient Primitives for Deep Learning*. 3rd Oct. 2014. arXiv: 1410.0759 [cs]. URL: http://arxiv.org/abs/1410.0759 (visited on 30/09/2019) (cit. on p. 89).

Chinchor, Nancy (1992). 'MUC-4 Evaluation Metrics'. In: 4th Message Understanding Conference. McLean, Virginia, 16th June 1992. URL: https://www.aclweb.org/anthology/M92-1002 (visited on 30/11/2019) (cit. on p. 20).

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (2014). 'Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2014. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: http://aclweb.org/anthology/D14-1179 (visited on 22/09/2019) (cit. on pp. 7, 71, 74, 79).

Choi, Heeyoul, Kyunghyun Cho and Yoshua Bengio (2017). 'Context-dependent word representation for neural machine translation'. In: *Computer Speech & Language* 45 (1st Sept. 2017), pp. 149–160. DOI: 10.1016/j.csl.2017.01.007. URL: http://www.sciencedirect.com/science/article/pii/S0885230816301024 (visited on 16/02/2020) (cit. on p. 101).

Chomsky, Noam (1981). *Lectures on Government and Binding: The Pisa Lectures*. Studies in Generative Grammar. Foris Publications (cit. on pp. 56, 106).

Clark, Kevin and Christopher D. Manning (2015). 'Entity-Centric Coreference Resolution with Model Stacking'. In: *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. ACL-IJCNLP 2015 (□□□, □□ (Beijing, China), 26th–31st July 2015). Vol. 1. Association for Computational Linguistics, pp. 1405–1415. DOI: 10.3115/v1/P15-1136. URL: http://aclweb.org/anthology/P15-1136 (visited on 16/09/2019) (cit. on pp. 53, 64).

Clark, Kevin and Christopher D. Manning (2016a). 'Deep Reinforcement Learning for Mention-Ranking Coreference Models'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2016 (Austin, Texas, USA, 1st–4th Nov. 2016), pp. 2256–2262. URL: http://aclweb.org/anthology/D/D16/D16-1245.pdf (cit. on pp. 53, 61, 84, 109).

Clark, Kevin and Christopher D. Manning (2016b). 'Improving Coreference Resolution by Learning Entity-Level Distributed Representations'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016 (Berlin, Deutschland, 7th–12th Aug. 2016). Vol. 1. Association for Computational Linguistics. URL: http://aclweb.org/anthology/P/P16/P16-1061.pdf (cit. on pp. 53–55, 64, 96, 98).

Clouzot, Catherine, Georges Antoniadis and Agnès Tutin (2000). 'Constitution and exploitation of an annotation system of electronic corpora: Toward automatic generation of understandable pronouns in French language'. In: *Natural Language Processing — NLP 2000*. Ed. by Dimitris N. Christodoulakis. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 242–251. DOI: 10.1007/3-540-45154-4_23 (cit. on p. 36).

Collobert, Ronan and Jason Weston (2008). 'A unified architecture for natural language processing: deep neural networks with multitask learning'. In: *Proceedings of the 25th international conference on Machine learning*. ICML 2008 (Helsinki, Suomi). ICML '08. Association for Computing Machinery, 5th July 2008, pp. 160–167. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177 (visited on 16/02/2020) (cit. on pp. 54, 99).

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel P. Kuksa (2011). 'Natural Language Processing (almost) from Scratch'. In: *Journal of Machine Learning Research* 12 (12th Aug. 2011), 2493–2537. URL: https://dl.acm.org/doi/10.5555/1953048.2078186 (visited on 05/11/2016) (cit. on pp. 52, 54).

Cornuéjols, Antoine and Laurent Miclet (2010). *Apprentissage artificiel - Concepts et algorithmes*. 2nd ed. Eyrolles. URL: https://www.eyrolles.com/Sciences/Livre/apprentissage-artificiel-9782212124712 (visited on 13/01/2018) (cit. on p. 7).

Crabbé, Benoit, Murielle Fabre and Christophe Pallier (2019). 'Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1150–1160. DOI:

10.18653/v1/D19-1106. URL: https://www.aclweb.org/anthology/D19-1106 (visited on 07/02/2020) (cit. on p. 112).

Cross, James and Liang Huang (2016). 'Span-Based Constituency Parsing with a Structure-Label System and Provably Optimal Dynamic Oracles'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2016 (Austin, Texas). Association for Computational Linguistics, pp. 1–11. DOI: 10.18653/v1/D16-1001. URL: http://aclweb.org/anthology/D16-1001 (visited on 02/03/2019) (cit. on pp. 4, 74).

Culotta, Aron, Michael Wick and Andrew McCallum (2007). 'First-Order Probabilistic Models for Coreference Resolution'. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. NAACL-HLT 2007. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 81–88. URL: https://www.aclweb.org/anthology/N07-1011 (visited on 29/09/2019) (cit. on pp. 52, 57, 85).

Dai, Andrew M and Quoc V Le (2015). 'Semi-supervised Sequence Learning'. In: *Advances in Neural Information Processing Systems*. NeurIPS 2015 (Montréal, Québec, Canada). Vol. 28. Curran Associates, Inc., pp. 3079–3087. URL: http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf (visited on 16/02/2020) (cit. on p. 101).

Delaborde, Marine and Frédéric Landragin (2019). 'En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on »'. In: *Les cahiers de praxématique*. La gestion de l'anaphore en discours : complexités et enjeux 72 (June 2019), pp. 1–18. URL: https://hal.archives-ouvertes.fr/hal-02161902 (visited on 15/10/2019) (cit. on pp. 12, 20).

Denis, Pascal and Jason Baldridge (2007a). 'A Ranking Approach to Pronoun Resolution'. In: *Proceddings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI 2007 (San Francisco, California, USA). Morgan Kaufmann, pp. 1588–1593. URL: http://dl.acm.org/citation.cfm?id=1625275.1625532 (visited on 19/07/2018) (cit. on p. 77).

Denis, Pascal and Jason Baldridge (2007b). 'Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming'. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. NAACL-HLT 2007 (Rochester, New York, USA). Association for Computational Linguistics, Apr. 2007, pp. 236–243. URL: https://www.aclweb.org/anthology/N07-1030 (visited on 26/09/2019) (cit. on pp. 78, 93).

Denis, Pascal and Jason Baldridge (2008). 'Specialized models and ranking for coreference resolution'. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2008 (Honolulu, Hawai'i, USA). Association for Computational Linguistics, p. 660. DOI: 10.3115/1613715.1613797. URL: http://portal.acm.org/citation.cfm?doid=1613715.1613797 (visited on 26/09/2019) (cit. on pp. 61, 77, 78).

Denis, Pascal and Jason Baldridge (2009). 'Global joint models for coreference resolution and named entity classification'. In: *Procesamiento del lenguaje natural* 42 (Mar. 2009). URL: http://rua.ua.es/dspace/handle/10045/10549 (visited on 22/08/2019) (cit. on pp. 25, 61, 93).

Désoyer, Adèle, Frédéric Landragin and Isabelle Tellier (2015a). 'Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC'. In: *Actes de la 22ème Conférence sur le Traitement Automatique des Langues*

Naturelles. TALN-RÉCITAL 2015 (Caen, France, 22nd–25th June 2015). June 2015, pp. 439–445. URL: https://halshs.archives-ouvertes.fr/halshs-01162174 (visited on 31/01/2017) (cit. on pp. 4, 65, 67, 77, 90).

Désoyer, Adèle, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre and Jean-Yves Antoine (2015b). 'Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR'. In: *Traitement Automatique des Langues*. Traitement automatique du langage parlé 55.2 (May 2015), pp. 97–121. URL: https://halshs.archives-ouvertes.fr/halshs-01153297 (visited on 31/01/2017) (cit. on pp. 39, 44).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 10th Oct. 2018. arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 16/02/2019) (cit. on pp. 4, 72, 73, 80, 81, 85, 99, 101, 102).

Dice, Lee R. (1945). 'Measures of the Amount of Ecologic Association Between Species'. In: *Ecology* 26.3, pp. 297–302. DOI: 10.2307/1932409. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409 (visited on 22/08/2019) (cit. on p. 24).

Diestel, Reinhard (2017). *Graph Theory*. 5th ed. Graduate Texts in Mathematics. Berlin Heidelberg: Springer-Verlag. DOI: 10.1007/978-3-662-53622-3. URL: https://www.springer.com/gp/book/9783662536216 (visited on 09/01/2020) (cit. on p. 7).

Dinarelli, Marco and Loïc Grobol (2018). 'Modélisation d'un contexte global d'étiquettes pour l'étiquetage de séquences dans les réseaux neuronaux récurrents'. In: *Actes de la onzième édition de la plate-forme Intelligence Artificielle*. Journée commune AFIA-ATALA sur le Traitement Automatique des Langues et l'Intelligence Artificielle pendant la onzième édition de la plate-forme Intelligence Artificielle. PFIA 2018. Nancy, France: As, 6th July 2018. URL: https://hal.archives-ouvertes.fr/hal-02002111/document (visited on 01/02/2019) (cit. on p. 71).

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel and Ralph Weischedel (2004). 'The Automatic Content Extraction (ACE) Program : Tasks, Data, and Evaluation'. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. LREC-2004. Lisbon, Portugal: European Language Resources Association, May 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf (visited on 06/07/2017) (cit. on pp. 14, 18, 30–32, 52, 64).

Dozat, Timothy (2016). 'Incorporating Nesterov Momentum into Adam'. In: Fourth International Conference on Learning Representations. San Juan, Puerto Rico, 2nd May 2016. URL: https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ (visited on 19/08/2019) (cit. on p. 80).

Dozat, Timothy and Christopher D. Manning (2017). 'Deep Biaffine Attention for Neural Dependency Parsing'. In: *5th International Conference on Learning Representations*. ICLR 2017. Toulon, France: OpenReview.net, 24th Apr. 2017. URL: https://openreview.net/forum?id=Hk95PK9le (visited on 24/07/2019) (cit. on p. 77).

Dupont, Yoann (2017). 'La structuration dans les entités nommées'. PhD Thesis. Université Sorbonne Paris Cité, 23rd Nov. 2017. URL: http://www.theses.fr/2017USPCA100 (visited on 06/08/2019) (cit. on pp. 16, 106).

Durrett, Greg, David Hall and Dan Klein (2013). 'Decentralized Entity-Level Modeling for Coreference Resolution'. In: Proceedings of the 51st Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers). Aug. 2013, pp. 114–124. URL: https://www.aclweb.org/anthology/P13-1012/ (visited on 04/08/2019) (cit. on p. 61).

Durrett, Greg and Dan Klein (2013). 'Easy Victories and Uphill Battles in Coreference Resolution'. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. Association for Computational Linguistics, pp. 1971–1982. URL: http://aclweb.org/anthology/D13-1203 (visited on 03/03/2019) (cit. on pp. 52–54, 60, 112).

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros and Noah A. Smith (2016). 'Recurrent Neural Network Grammars'. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL: HLT 2016. NAACL: HLT 2016. San Diego, California, USA: Association for Computational Linguistics, pp. 199–209. DOI: 10.18653/v1/N16-1024. URL: http://aclweb.org/anthology/N16-1024 (visited on 12/03/2019) (cit. on p. 112).

Edmonds, Jack (1967). 'Optimum branchings'. In: *Journal of Research of the national Bureau of Standards B* 71.4, pp. 233–240 (cit. on p. 62).

Eger, Steffen, Paul Youssef and Iryna Gurevych (2018). 'Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP tasks'. In: 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 4415–4424. DOI: 10.18653/v1/D18-1472. URL: http://aclweb.org/anthology/D18-1472 (visited on 30/09/2019) (cit. on p. 88).

Ehrmann, Maud (2008). 'Named entities, from Linguistics to NLP: Theoretical status and disambiguation methods'. Theses. Université Paris Diderot, June 2008. URL: https://hal.archives-ouvertes.fr/tel-01639190 (visited on 21/11/2019) (cit. on pp. 15, 16).

Eshkol-Taravella, Iris, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua and Isabelle Tellier (2011). 'Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012'. In: *Traitement Automatique des Langues*. Ressources Linguistiques Libres 53.2, pp. 17–46. URL: https://halshs.archives-ouvertes.fr/halshs-01163053 (visited on 08/01/2020) (cit. on p. 36).

Evert, Stefan and Andrew Hardie (2011). 'Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium'. In: *Proceedings of the Corpus Linguistics 2011 conference*. CL2011. Birmingham, United Kingdom (cit. on p. 47).

Fauconnier, Gilles, George Lakoff and Eve Sweester (1994). *Mental spaces: aspects of meaning construction in natural language*. 2nd ed. Cambridge University Press (cit. on p. 9).

Fernandes, Eraldo, Cícero dos Santos and Ruy Milidiú (2012). 'Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution'. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012 (□□□, □□□□ (Jeju Island, Republic of Korea)). Association for Computational Linguistics, July 2012, pp. 41–48. URL: https://www.aclweb.org/anthology/W12-4502 (visited on 09/02/2020) (cit. on p. 62).

Fernandes, Eraldo R. and Ulf Brefeld (2011). 'Learning from Partially Annotated Sequences'. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba and Michalis Vazirgiannis. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 407–422. DOI: 10.1007/978-3-642-23780-5_36 (cit. on p. 62).

Fernandes, Eraldo Rezende, Cícero Nogueira dos Santos and Ruy Luiz Milidiú (2014). 'Latent Trees for Coreference Resolution'. In: *Computational Linguistics* 40.4 (Dec. 2014), pp. 801–835. DOI: 10.1162/COLI_a_00200. URL: https://www.aclweb.org/anthology/J14-4004 (visited on 09/02/2020) (cit. on p. 62).

Finkel, Jenny Rose and Christopher D. Manning (2008). 'Enforcing Transitivity in Coreference Resolution'. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. HLT 2008 (Columbus, Ohio, USA). Association for Computational Linguistics, June 2008, pp. 45–48. URL: https://www.aclweb.org/anthology/P08-2012 (visited on 30/11/2019) (cit. on p. 23).

Finkel, Jenny Rose and Christopher D. Manning (2009). 'Joint Parsing and Named Entity Recognition'. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLR: NAACL 2009 (Boulder, Colorado, USA). Association for Computational Linguistics, pp. 326–334. URL: http://dl.acm.org/citation.cfm?id=1620754.1620802 (visited on 26/03/2019) (cit. on p. 33).

Flaubert, Gustave (1910). *Salammbô*. Paris, France: Louis Conard. URL: https://fr.wikisource.org/wiki/Salammb%C3%B4 (cit. on p. 22).

Fort, Karën, Adeline Nazarenko and Ris Claire (2011). 'Corpus Linguistics for the Annotation Manager'. In: Corpus Linguistics 2011. Birmingham, United Kingdom, 20th July 2011. URL: https://hal.archives-ouvertes.fr/hal-00641571 (visited on 10/12/2019) (cit. on p. 30).

Fort, Karën, Adeline Nazarenko and Sophie Rosset (2012). 'Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis'. In: *Proceedings of COLING 2012*. COLING 2012. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 895–910. URL: https://www.aclweb.org/anthology/C12-1055 (visited on 20/12/2019) (cit. on p. 31).

French, Robert Matthew (1999). 'Catastrophic forgetting in connectionist networks'. In: *Trends in Cognitive Sciences* 3.4 (Apr. 1999), pp. 128–135. pmid: 10322466 (cit. on p. 85).

Gaier, Adam and David Ha (2019). *Weight Agnostic Neural Networks*. 10th June 2019. arXiv: 1906.04358 [cs, stat]. URL: http://arxiv.org/abs/1906.04358 (visited on 28/09/2019) (cit. on p. 80).

Gal, Yarin and Zoubin Ghahramani (2015). *A Theoretically Grounded Application of Dropout in Recurrent Neural Networks*. 16th Dec. 2015. arXiv: 1512.05287 [stat]. URL: http://arxiv.org/abs/1512.05287 (visited on 21/02/2019) (cit. on p. 81).

Galliano, Sylvain, Guillaume Gravier and Laura Chaubard (2009). 'The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts'. In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association*. InterSpeech 2009 (Brighton, United Kingdom). International Speech Communication Association (cit. on p. 106).

Gardent, Claire and Hélène Manuélian (2005). 'Création d'un corpus annoté pour le traitement des descriptions définies'. In: *Traitement Automatique des Langues*. Modèles et algorithmes pour la résolution d'anaphores 1.46, pp. 115–140. URL: https://members.loria.fr/CGardent/publis/tal05-dede.pdf (visited on 07/01/2020) (cit. on p. 36).

Gillick, Daniel, Benoit Favre, Dilek Hakkani-Tür, Berndt Bohnet, Yang Liu and Shasha Xie (2009). 'The ICSI/UTD Summarization System at TAC 2009'. In: *Proceedings of the Text Analysis*

Conference workshop. TAC 2009 (Gaithersburg, Maryland, USA). URL: https://hal-amu.archives-ouvertes.fr/hal-01194277 (visited on 20/01/2020) (cit. on p. 33).

Godbert, Elisabeth and Benoît Favre (2017). 'Détection de coréférences de bout en bout en français'. In: *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2017. Orléans, France: Association pour le Traitement Automatique des Langues, June 2017. URL: https://hal.archives-ouvertes.fr/hal-01687116 (cit. on pp. 65, 67).

Godin, Frédéric (2019). 'Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing'. PhD Thesis. Gent, België: Universiteit Gent, July 2019. 214 pp. URL: https://fredericgodin.com/wp-content/uploads/2019/07/phd_frederic_godin_book.pdf (visited on 17/02/2020) (cit. on p. 103).

Goldberg, Yoav and Joakim Nivre (2012). 'A Dynamic Oracle for Arc-Eager Dependency Parsing'. In: *Proceedings of the 24th International Conference on Computational Linguistics*. COLING 2012 (Mumbai, India). Dec. 2012, pp. 959–976. URL: https://www.aclweb.org/anthology/C12-1059 (visited on 10/02/2020) (cit. on p. 63).

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomáš Mikolov (2018). 'Learning Word Vectors for 157 Languages'. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*. LREC 2018 (Miyazaki, Japan). European Language Resource Association. URL: http://aclweb.org/anthology/L18-1550 (visited on 03/03/2019) (cit. on pp. 87, 99, 100).

Graves, Alex and Jürgen Schmidhuber (2005). 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures'. In: *Neural Networks* 18.5-6, pp. 602–610. DOI: 10.1016/j.neunet.2005.06.042. pmid: 16112549 (cit. on pp. 6, 73).

Gravier, Guillaume, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel and Olivier Galibert (2012). 'The ETAPE corpus for the evaluation of speech-based TV content processing in the French language'. In: *LREC - Eighth international conference on Language Resources and Evaluation*. Turkey, na. URL: https://hal.archives-ouvertes.fr/hal-00712591 (visited on 18/02/2020) (cit. on p. 106).

Grice, Herbert Paul (1989). *Studies in the Way of Words*. Harvard University Press. 406 pp. (cit. on p. 10).

Grobol, Loïc (2019). 'Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French'. In: *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*. CRAC 2019 (Minneapolis, Minnesota). June 2019, pp. 8–14. URL: https://www.aclweb.org/anthology/papers/W/W19/W19-2802/ (visited on 24/06/2019) (cit. on pp. 86, 89, 90).

Grobol, Loïc, Frédéric Landragin and Serge Heiden (2017a). 'Interoperable annotation of (co)references in the Democrat project'. In: *Proceedings of the thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation*. ISA13 (Montpellier, France). Ed. by Harry Bunt. ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2, Sept. 2017. URL: https://hal.archives-ouvertes.fr/hal-01583527 (cit. on p. 41).

Grobol, Loïc, Frédéric Landragin and Serge Heiden (2018a). 'XML-TEI-URS: using a TEI format for annotated linguistic resources'. In: CLARIN Annual Conference 2018 (Pisa, Italia). Oct.

2018. URL: https://hal.archives-ouvertes.fr/hal-01827563 (visited on 30/01/2019) (cit. on p. 41).

Grobol, Loïc, Isabelle Tellier, Éric Villemonte De La Clergerie, Marco Dinarelli and Frédéric Landragin (2017b). 'Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral'. In: *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2017 (Orléans, France). Association pour le Traitement Automatique des Langues, June 2017. URL: https://hal.inria.fr/hal-01558711 (cit. on pp. 68, 89).

Grobol, Loïc, Isabelle Tellier, Éric Villemonte De La Clergerie, Marco Dinarelli and Frédéric Landragin (2018b). 'ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations'. In: *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*. LREC 2018 (Miyazaki, Japan). European Language Resource Association, May 2018. URL: https://hal.inria.fr/hal-01744572 (cit. on pp. 41, 45, 107).

Grönroos, Stig-Arne, Sami Virpioja, Peter Smit and Mikko Kurimo (2014). 'Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology'. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. COLING 2014 (Baile Átha Cliath, Éire). Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1177–1185. URL: https://www.aclweb.org/anthology/C14-1111 (visited on 16/02/2020) (cit. on p. 99).

Habert, Benoît (2000). 'Détournements d'annotation : armer la main et le regard'. In: *Corpus: méthodologie et applications linguistiques*. Ed. by Mireille Bilger. 1 vols. Paris, France: Honoré Champion, pp. 106–120 (cit. on p. 30).

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal and James R. Curran (2013). 'Evaluating Entity Linking with Wikipedia'. In: *Artificial Intelligence*. Artificial Intelligence, Wikipedia and Semi-Structured Resources 194 (1st Jan. 2013), pp. 130–150. DOI: 10.1016/j.artint.2012.04.005. URL: http://www.sciencedirect.com/science/article/pii/S0004370212000446 (visited on 30/11/2019) (cit. on p. 16).

Haghighi, Aria and Dan Klein (2009). 'Simple Coreference Resolution with Rich Syntactic and Semantic Features'. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2009 (Singapore). Association for Computational Linguistics, Aug. 2009, pp. 1152–1161. URL: https://www.aclweb.org/anthology/D09-1120 (visited on 30/01/2020) (cit. on p. 52).

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue and Gong Bing (2017). 'Learning from class-imbalanced data: Review of methods and applications'. In: *Expert Systems with Applications* 73 (1st May 2017), pp. 220–239. DOI: 10.1016/j.eswa.2016.12.035. URL: http://www.sciencedirect.com/science/article/pii/S0957417416307175 (visited on 21/09/2019) (cit. on p. 83).

Hanson, Stephen José and Lorien Y. Pratt (1988). 'Comparing Biases for Minimal Network Construction with Back-propagation'. In: *Proceedings of the 1st International Conference on Neural Information Processing Systems*. NIPS'88. Cambridge, MA, USA: MIT Press, pp. 177–185. URL: http://dl.acm.org/citation.cfm?id=2969735.2969756 (visited on 06/09/2019) (cit. on p. 82).

Hatmi, Mohamed (2014). 'Named entity recognition in multimodal documents'. Theses. Université de Nantes, Jan. 2014. URL: https://hal.archives-ouvertes.fr/tel-01154811 (visited on 18/02/2020) (cit. on p. 106).

Heiden, Serge (2019). *Manuel de TXM, Extension Annotation URS (Unité-Relation-Schéma) version 1.0*. Manual. 3rd July 2019. DOI: 10.5281/zenodo.3267345. URL: https://zenodo.org/record/3267345 (visited on 23/01/2020) (cit. on p. 47).

Heiden, Serge, Jean-Philippe Magué and Bénédicte Pincemin (2010). 'TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement'. In: *10th International Conference on the Statistical Analysis of Textual Data*. JADT 2010. Ed. by Luca Giuliano Sergio Bolasco Isabella Chiari. Vol. 2. Roma, Italia: Edizioni Universitarie di Lettere Economia Diritto, May 2010, pp. 1021–1032. URL: https://halshs.archives-ouvertes.fr/halshs-00549779 (visited on 06/07/2017) (cit. on pp. 40, 46).

Heinzerling, Benjamin and Michael Strube (2019). 'Sequence Tagging with Contextual and Non-Contextual Subword Representations: A Multilingual Evaluation'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019 (Firenze, Italia). Florence, Italy: Association for Computational Linguistics, July 2019, pp. 273–291. DOI: 10.18653/v1/P19-1027. URL: https://www.aclweb.org/anthology/P19-1027 (visited on 17/02/2020) (cit. on p. 103).

Hendrickx, Iris, Gosse Bouma, Walter Daelemans and Véronique Hoste (2013). 'COREA: Coreference Resolution for Extracting Answers for Dutch'. In: *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*. Ed. by Peter Spyns and Jan Odijk. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 115–128. DOI: 10.1007/978-3-642-30910-6_7. URL: https://doi.org/10.1007/978-3-642-30910-6_7 (visited on 19/01/2020) (cit. on p. 33).

Hinrichs, Erhard, Sandra Kübler, Karin Naumann, Heike Telljohann and Julia Trushkina (2004). 'Recent developments in linguistic annotations of the TüBa-D/Z treebank'. In: *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. TLT 2004 (10th–11th Dec. 2004). Ed. by Sarah Kübler, Joakim Nivre, Erhard Hinrichs and Wolger Wunsch. Tübingen, Deutschland, pp. 51–62 (cit. on p. 32).

Hinton, Geoffrey E. (1987). 'Learning translation invariant recognition in a massively parallel networks'. In: *PARLE Parallel Architectures and Languages Europe*. Ed. by J. W. de Bakker, A. J. Nijman and P. C. Treleaven. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–13 (cit. on p. 82).

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever and Ruslan R. Salakhutdinov (2012). *Improving neural networks by preventing co-adaptation of feature detectors*. 3rd July 2012. arXiv: 1207.0580 [cs]. URL: http://arxiv.org/abs/1207.0580 (visited on 06/09/2019) (cit. on p. 81).

Hirschman, Lynette and Nancy Chinchor (1998). 'Appendix F: MUC-7 Coreference Task Definition (version 3.0)'. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference. Fairfax, Virginia, 29th Apr. 1998. URL: https://www.aclweb.org/anthology/M98-1029 (visited on 04/08/2019) (cit. on pp. 21, 64).

Hirst, G. (1981). *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science. Berlin Heidelberg: Springer-Verlag. DOI: 10.1007/3-540-10858-0. URL: https://www.springer.com/fr/book/9783540108580 (visited on 24/11/2019) (cit. on p. 19).

Hobbs, Jerry R. (1986). 'Resolving Pronoun References'. In: *Readings in Natural Language Processing*. Ed. by Barbara J. Grosz, Karen Sparck-Jones and Bonnie Lynn Webber. San Francisco, California, USA: Morgan Kaufmann, pp. 339–352. URL: http://dl.acm.org/citation.cfm?id=21922.24343 (cit. on pp. 2, 29, 52).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8 (1st Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735 (visited on 13/09/2019) (cit. on pp. 6, 54, 79).

Honnibal, Matthew and Ines Montani (2019). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Version 2.2.5. Explosion. URL: https://spacy.io (cit. on p. 103).

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel (2006). 'OntoNotes: The 90% Solution'. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Compuational Linguistics*. HLT NAACL 06. Vol. Short papers. New York City, USA: Association for Computational Linguistics, June 2006, pp. 57–60. URL: https://www.aclweb.org/anthology/N06-2015 (visited on 18/12/2019) (cit. on pp. 5, 14, 30–33, 41).

Ioffe, Sergey and Christian Szegedy (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 10th Feb. 2015. arXiv: 1502.03167 [cs]. URL: http://arxiv.org/abs/1502.03167 (visited on 10/09/2019) (cit. on pp. 81, 82).

Isard, Amy, David McKelvie, Andreas Mengel and Morten Baun Møller (2000). 'The MATE Workbench: A Tool for Annotating XML Corpora'. In: *Content-Based Multimedia Information Access*. RIAO 00 (Paris, France). Vol. 1. Centre de hautes études internationales d'informatique documentaire, pp. 411–425. URL: http://dl.acm.org/citation.cfm?id=2835865.2835908 (cit. on p. 40).

ISO/TC 37/SC 4/WG 2 (2017). *ISO AWI 24617-9 Language resource management – Part 9 Semantic annotation framework (SemAF)*. Reference. Geneva, Switzerland: International Organization for Standardization. URL: https://www.iso.org/standard/69658.html (cit. on p. 44).

Jaccard, Paul (1912). 'The Distribution of the Flora in the Alpine Zone'. In: *New Phytologist* 11.2, pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x (visited on 16/02/2020) (cit. on p. 98).

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer and Omer Levy (2019a). *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. 24th July 2019. arXiv: 1907.10529 [cs]. URL: http://arxiv.org/abs/1907.10529 (visited on 25/07/2019) (cit. on pp. 54, 55, 61, 73).

Joshi, Mandar, Omer Levy, Daniel S. Weld and Luke Zettlemoyer (2019b). *BERT for Coreference Resolution: Baselines and Analysis*. 24th Aug. 2019. arXiv: 1908.09091 [cs]. URL: http://arxiv.org/abs/1908.09091 (visited on 01/09/2019) (cit. on pp. 54, 55, 61, 73, 101, 102).

Jouppi, Norman P. et al. (2017). 'In-Datacenter Performance Analysis of a Tensor Processing Unit'. In: *ACM SIGARCH Computer Architecture News* 45.2 (24th June 2017), pp. 1–12. DOI: 10.1145/3140659.3080246. URL: https://doi.org/10.1145/3140659.3080246 (visited on 16/02/2020) (cit. on p. 101).

Jurafsky, Daniel and James H. Martin (2019). *Speech and Language Processing*. 3rd ed. Pearson. 1032 pp. URL: https://web.stanford.edu/~jurafsky/slp3/ (cit. on p. 7).

Kantor, Ben and Amir Globerson (2019). 'Coreference Resolution with Entity Equalization'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019 (Firenze, Italia). Association for Computational Linguistics, pp. 673–677. URL: https://www.aclweb.org/anthology/P19-1066/ (visited on 31/08/2019) (cit. on pp. 50, 54, 55).

Karpathy, Andrej and Justin Johnson (2019). 'CS231n Convolutional Neural Networks for Visual Recognition'. Lecture notes. Lecture notes. URL: http://cs231n.github.io/ (visited on 27/09/2019) (cit. on pp. 80, 91).

Karttunen, Lauri (1976). 'Discourse Referents'. In: *Notes from the Linguistic Underground*. Ed. by James David McCawley. Vol. 7. Syntax and Semantics. Academic Press (cit. on pp. 8, 9, 112).

Kent, Allen, Madeline M. Berry, Fred U. Jr. Luehrs and J. W. Perry (1955). 'Machine literature searching VIII. Operational criteria for designing information retrieval systems'. In: *American Documentation* 6.2, pp. 93–101. DOI: 10.1002/asi.5090060209. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090060209 (visited on 30/11/2019) (cit. on p. 20).

Kingma, Diederik P. and Jimmy Ba (2014). 'Adam: A Method for Stochastic Optimization'. In: International Conference on Learning Representations. San Diego, California, 22nd Dec. 2014. arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on 04/03/2019) (cit. on p. 80).

Kingma, Diederik P., Tim Salimans and Max Welling (2015). *Variational Dropout and the Local Reparameterization Trick*. 8th June 2015. arXiv: 1506.02557 [cs, stat]. URL: http://arxiv.org/abs/1506.02557 (visited on 06/09/2019) (cit. on p. 81).

Klenner, Manfred (2007). 'Enforcing consistency on coreference sets'. In: Proceedings of Recent Advances in Natural Processing 2007 (Боровец, България). RANLP 2007, pp. 323–328. URL: https://www.zora.uzh.ch/id/eprint/19142/1/ranlp07.pdf (cit. on p. 61).

Kopeć, Mateusz and Maciej Ogrodniczuk (2012). 'Creating a Coreference Resolution System for Polish'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. LREC 2012. European Language Resources Association, May 2012, pp. 192–195. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1064_Paper.pdf (visited on 31/01/2020) (cit. on p. 52).

Krogh, Anders and John A. Hertz (1991). 'A Simple Weight Decay Can Improve Generalization'. In: 4th International Conference on Neural Information Processing Systems (Denver, Colorado). NIPS'91. San Francisco, California: Morgan Kaufmann, pp. 950–957. URL: http://dl.acm.org/citation.cfm?id=2986916.2987033 (visited on 06/09/2019) (cit. on p. 82).

Krueger, David et al. (2016). *Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations*. 3rd June 2016. arXiv: 1606.01305 [cs]. URL: http://arxiv.org/abs/1606.01305 (visited on 06/09/2019) (cit. on p. 81).

Kruskal, Joseph B. (1956). 'On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem'. In: *Proceedings of the American Mathematical Society* 7.1, pp. 48–50. DOI: 10.2307/2033241. JSTOR: 2033241 (cit. on p. 62).

Kuhn, Harold (1955). 'The Hungarian method for the assignment problem'. In: *Naval Research Logistics Quarterly* 2, pp. 83–97 (cit. on p. 24).

Kummerfeld, Jonathan K., Mohit Bansal, David Burkett and Dan Klein (2011). 'Mention Detection: Heuristics for the OntoNotes Annotations'. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. CoNLL 2011 (Portland, Oregon, USA). CONLL Shared Task '11. Association for Computational Linguistics, pp. 102–106. URL: http://dl.acm.org/citation.cfm?id=2132936.2132953 (cit. on pp. 32, 34).

Lacheret, Anne et al. (2014). 'Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. LREC 2014 (tex.venue: Reykjavik, Ísland). European Language Resource Association, May 2014. URL: http://hal.upmc.fr/hal-00968959 (cit. on pp. 46, 75).

Lafferty, John, Andrew McCallum and Fernando Pereira (2001). 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. In: 18th International Conference on Machine Learning. ICML '01. San Francisco, CA, USA: Morgan Kaufmann, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813 (visited on 30/04/2019) (cit. on p. 16).

Landragin, Frédéric (2016). 'Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)'. In: *Bulletin de l'AFIA* 92, pp. 11–15 (cit. on pp. 5, 35, 111).

Landragin, Frédéric, Marine Delaborde, Yoann Dupont and Loïc Grobol (2018). *Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours*. May 2018. URL: https://hal.archives-ouvertes.fr/hal-01797982 (visited on 30/01/2019) (cit. on p. 40).

Landragin, Frédéric, Thierry Poibeau and Bernard Victorri (2012). 'ANALEC: a New Tool for the Dynamic Annotation of Textual Data'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. LREC 2012 (İstanbul, Türkiye). European Language Resources Association, May 2012, pp. 357–362. URL: https://halshs.archives-ouvertes.fr/halshs-00698971 (visited on 07/07/2017) (cit. on pp. 41, 43).

Lassalle, Emmanuel (2015). 'Structured learning with latent trees: a joint approach to coreference resolution'. PhD Thesis. Université Paris Diderot Paris 7, May 2015. URL: https://hal.inria.fr/tel-01331425 (visited on 03/10/2016) (cit. on pp. 59, 62).

Lassalle, Emmanuel and Pascal Denis (2015). 'Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures'. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI 2015. Austin, Texas, 25th Jan. 2015. URL: https://hal.inria.fr/hal-01205189/document (visited on 26/03/2019) (cit. on pp. 62, 78).

Le, Hang et al. (2020). 'FlauBERT: Unsupervised Language Model Pre-training for French'. In: *Proceedings of the 13th Language Resources and Evaluation Conference*. LREC 2020 (Marseille, France). European Language Resource Association. URL: https://hal.archives-ouvertes.fr/hal-02890258 (visited on 27/07/2020) (cit. on p. 101).

Le, Phong and Ivan Titov (2017). 'Optimizing Differentiable Relaxations of Coreference Evaluation Metrics'. In: 21st Conference on Computational Natural Language Learning. Vancouver,

Canada: Association for Computational Linguistics, pp. 390–399. DOI: 10.18653/v1/K17-1039. URL: http://aclweb.org/anthology/K17-1039 (visited on 01/09/2019) (cit. on pp. 55, 84).

LeCun, Yann (1988). 'A theoretical framework for back-propagation'. In: *Proceedings of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA*, pp. 21–28. URL: https://nyuscholars.nyu.edu/en/publications/a-theoretical-framework-for-back-propagation-2 (visited on 17/09/2019) (cit. on p. 80).

LeCun, Yann, Léon Bottou, Genevieve B. Orr and Klaus-Robert Müller (2012). 'Efficient BackProp'. In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr and Klaus-Robert Müller. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–48. DOI: 10.1007/978-3-642-35289-8_3. URL: https://doi.org/10.1007/978-3-642-35289-8_3 (visited on 28/09/2019) (cit. on p. 80).

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky (2013). 'Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules'. In: *Computational Linguistics* 39.4 (Dec. 2013), pp. 885–916. DOI: 10.1162/COLI_a_00152. URL: http://dx.doi.org/10.1162/COLI_a_00152 (visited on 11/10/2016) (cit. on pp. 52, 63, 64).

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*. CoNLL 2011 (Portland, Oregon, USA). CONLL Shared Task '11. Association for Computational Linguistics, June 2011, pp. 28–34. URL: https://www.aclweb.org/anthology/W11-1902 (visited on 31/01/2020) (cit. on pp. 52, 63–65).

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu and Dan Jurafsky (2012). 'Joint Entity and Event Coreference Resolution Across Documents'. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012 (□□□, □□□□ (Jeju Island, Republic of Korea)). Association for Computational Linguistics, pp. 489–500. URL: http://dl.acm.org/citation.cfm?id=2390948.2391006 (visited on 24/11/2019) (cit. on p. 19).

Lee, Heeyoung, Mihai Surdeanu and Dan Jurafsky (2017). 'A scaffolding approach to coreference resolution integrating statistical and rule-based models'. In: *Natural Language Engineering* 23.5 (Sept. 2017), pp. 733–762. DOI: 10.1017/S1351324917000109. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/scaffolding-approach-to-coreference-resolution-integrating-statistical-and-rulebased-models/042D0D6C6E125EFB939E0F2C2E63152B (visited on 31/01/2020) (cit. on pp. 63, 64, 98).

Lee, Kenton, Luheng He, Mike Lewis and Luke Zettlemoyer (2017). 'End-to-end Neural Coreference Resolution'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2017. København, Danmark: Association for Computational Linguistics, Sept. 2017, pp. 188–197. URL: https://www.aclweb.org/anthology/D17-1018 (cit. on pp. 4, 50, 53–55, 61, 64, 68, 69, 74, 75, 78–80, 84, 85, 89, 93, 96, 101, 109).

Lee, Kenton, Luheng He and Luke Zettlemoyer (2018). 'Higher-Order Coreference Resolution with Coarse-to-Fine Inference'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

NAACL: HLT 2018 (New Orleans, Louisiana). Vol. 2. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: http://aclweb.org/anthology/N18-2108 (visited on 16/02/2019) (cit. on pp. 4, 50, 54, 55, 72, 77, 79, 84, 85, 89, 101).

Leech, Geoffrey (2005). 'Adding Linguistic Annotation'. In: Wynne, Martin. *Developing linguistic corpora: A guide to good practice*. AHDS Guides to Good Practice. Oxford: Oxbow Books, 16th Sept. 2005, pp. 25–39. URL: https://web.archive.org/web/20160819223836/http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm (cit. on pp. 29, 30).

Levesque, Hector, Ernest Davis and Leora Morgenstern (2012). 'The Winograd Schema Challenge'. In: Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (Rome, Italy). KR'12. AAAI Press, pp. 552–561. URL: http://dl.acm.org/citation.cfm?id=3031843.3031909 (visited on 30/11/2019) (cit. on p. 18).

Levy, Omer, Kenton Lee, Nicholas FitzGerald and Luke Zettlemoyer (2018). *Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum*. 9th May 2018. arXiv: 1805.03716 [cs, stat]. URL: http://arxiv.org/abs/1805.03716 (visited on 08/02/2019) (cit. on p. 73).

Lin, Zhouhan, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou and Yoshua Bengio (2017). 'A Structured Self-Attentive Sentence Embedding'. In: *Proceedings of the 5th International Conference on Learning Representations*. 5th International Conference on Learning Representations. Toulon, France, 24th Apr. 2017. URL: https://openreview.net/forum?id=BJC_jUqxe (visited on 22/07/2019) (cit. on p. 75).

Ling, Wang, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo and Tiago Luis (2015). 'Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 1520–1530. DOI: 10.18653/v1/D15-1176. URL: http://aclweb.org/anthology/D15-1176 (visited on 03/09/2019) (cit. on pp. 4, 71).

Liu, Fei, Luke Zettlemoyer and Jacob Eisenstein (2019). 'The Referential Reader: A Recurrent Entity Network for Anaphora Resolution'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019 (Firenze, Italia). Association for Computational Linguistics, July 2019, pp. 5918–5925. DOI: 10.18653/v1/P19-1593. URL: https://www.aclweb.org/anthology/P19-1593 (visited on 06/02/2020) (cit. on pp. 57, 63, 94, 112).

Liu, Frederick, Han Lu and Graham Neubig (2018). 'Handling Homographs in Neural Machine Translation'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2018 (New Orleans, Louisiana). Vol. 1. Association for Computational Linguistics, June 2018, pp. 1336–1345. DOI: 10.18653/v1/N18-1121. URL: https://www.aclweb.org/anthology/N18-1121 (visited on 16/02/2020) (cit. on p. 101).

Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao and Jiawei Han (2019). *On the Variance of the Adaptive Learning Rate and Beyond*. 8th Aug. 2019. arXiv: 1908.03265 [cs, stat]. URL: http://arxiv.org/abs/1908.03265 (visited on 17/08/2019) (cit. on p. 80).

Liu, Yinhan et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 26th July 2019. arXiv: 1907.11692 [cs]. URL: http://arxiv.org/abs/1907.11692 (visited on 05/09/2019) (cit. on p. 81).

Lonergan, Joanna, Jack Kay and John Ross (1974). *Étude sociolinguistique sur Orléans : catalogue des enregistrements*. Catalog. Orléans, France: Orléans archive. URL: https://archivesetmanuscrits.bnf.fr/ark:/12148/cc95934w/ca19843673 (cit. on p. 37).

Longo, Laurence (2013). 'Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence'. PhD Thesis. Université de Strasbourg, Dec. 2013. URL: https://tel.archives-ouvertes.fr/tel-00939243 (visited on 31/01/2017) (cit. on pp. 65, 67).

Loshchilov, Ilya and Frank Hutter (2019). 'Decoupled Weight Decay Regularization'. In: 7th International Conference on Learning Representations. New Orleans, Louisiana. arXiv: 1711.05101. URL: https://arxiv.org/abs/1711.05101 (visited on 03/03/2019) (cit. on pp. 80, 82).

Lu, Jing and Vincent Ng (2018). 'Event Coreference Resolution: A Survey of Two Decades of Research'. In: *International Joint Conferences on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence. Stockholm, Sverige, 13th July 2018, pp. 5479–5486. URL: https://www.ijcai.org/proceedings/2018/773 (visited on 23/11/2019) (cit. on p. 18).

Luo, Xiaoqiang (2005). 'On Coreference Resolution Performance Metrics'. In: *Proceedingss of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05 (Vancouver, British Columbia, Canada). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 25–32. DOI: 10.3115/1220575.1220579. URL: https://doi.org/10.3115/1220575.1220579 (visited on 22/08/2019) (cit. on p. 24).

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla and Salim Roukos (2004). 'A Mention-Synchronous Coreference Resolution Algorithm Based On the Bell Tree'. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. ACL 2004 (Barcelona, España). Association for Computational Linguistics, July 2004, pp. 135–142. DOI: 10.3115/1218955.1218973. URL: https://www.aclweb.org/anthology/P04-1018 (visited on 30/11/2019) (cit. on pp. 52, 56, 63).

Luo, Xiaoqiang, Sameer Pradhan, Marta Recasens and Eduard Hovy (2014). 'An Extension of BLANC to System Mentions'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL 2014 (Baltimore, Maryland, USA), pp. 24–29. URL: http://69.195.124.161/%20aclwebor/anthology/P/P14/P14-2005.pdf (visited on 12/04/2017) (cit. on pp. 25, 26).

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts (2011). 'Learning Word Vectors for Sentiment Analysis'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL-HLT 2011 (Portland, Oregon, USA). Association for Computational Linguistics, June 2011, pp. 142–150. URL: https://www.aclweb.org/anthology/P11-1015 (visited on 16/02/2020) (cit. on p. 99).

Manjavacas, Enrique, Ákos Kádár and Mike Kestemont (2019). 'Improving Lemmatization of Non-Standard Languages with Joint Learning'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2019 (Minneapolis, Minnesota). Vol. 1. Association for

Computational Linguistics, June 2019, pp. 1493–1503. DOI: 10.18653/v1/N19-1153. URL: https://www.aclweb.org/anthology/N19-1153 (visited on 17/02/2020) (cit. on p. 103).

Martens, James (2016). 'Second-order Optimization for Neural Networks'. PhD Thesis. Toronto, Canada: University of Toronto. 179 pp. URL: http://www.cs.toronto.edu/~jmartens/docs/thesis_phd_martens.pdf (visited on 28/09/2019) (cit. on p. 80).

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah and Benoît Sagot (2020). 'CamemBERT: a Tasty French Language Model'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020 (Online). Association for Computational Linguistics, July 2020, pp. 7203–7219. URL: https://www.aclweb.org/anthology/2020.acl-main.645 (visited on 09/07/2020) (cit. on p. 101).

Martschat, Sebastian, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt and Michael Strube (2012). 'A Multigraph Model for Coreference Resolution'. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL 2012 (□□□, □□□□ (Jeju Island, Republic of Korea)). Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 100–106. URL: https://www.aclweb.org/anthology/W12-4511 (visited on 09/02/2020) (cit. on p. 59).

Mathet, Yann and Antoine Widlöcher (2011). 'Stratégie d'exploration de corpus multi-annotés avec GlozzQL'. In: *Actes de la 18e Conférence Traitement Automatique des Langues Naturelles*. TALN 2011 (Montpellier, France). Ed. by Lafourcade, Mathieu and Prince, Violaine. Vol. 1. Association pour le Traitement Automatique des Langues, June 2011, pp. 143–148. URL: https://hal.archives-ouvertes.fr/hal-01021846 (visited on 23/01/2020) (cit. on p. 47).

McCallum, Andrew and Ben Wellner (2004). 'Conditional Models of Identity Uncertainty with Application to Noun Coreference'. In: *Advances in Neural Information Processing Systems*. 17th International Conference on Neural Information Processing Systems (NeurIPS 2004) (Vancouver, British Columbia, Canada). Ed. by Lawrence Saul K, Yair Weiss and Léon Bottou. Vol. 17. Cambridge, MA, USA: MIT Press, pp. 905–912. URL: http://dl.acm.org/citation.cfm?id=2976040.2976154 (visited on 29/09/2019) (cit. on p. 78).

McCarthy, Joseph F. and Wendy G. Lehnert (1995). 'Using decision trees for conference resolution'. In: *Proceedings of the 14th international joint conference on Artificial intelligence*. IJCAI 1995. Vol. 2. Montréal, Québec, Canada: Morgan Kaufmann Publishers Inc., 20th Aug. 1995, pp. 1050–1055 (cit. on pp. 51, 59, 60, 98).

McCoy, R. Thomas, Ellie Pavlick and Tal Linzen (2019). *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*. 3rd Feb. 2019. arXiv: 1902.01007 [cs]. URL: http://arxiv.org/abs/1902.01007 (visited on 27/06/2019) (cit. on p. 72).

Meechan-Maddon, Ailsa and Joakim Nivre (2019). 'How to Parse Low-Resource Languages: Cross-Lingual Parsing, Target Language Annotation, or Both?' In: *Proceedings of the Fifth International Conference on Dependency Linguistics*. Depling, SyntaxFest 2019 (Paris, France). Association for Computational Linguistics, Aug. 2019, pp. 112–120. DOI: 10.18653/v1/W19-7713. URL: https://www.aclweb.org/anthology/W19-7713 (visited on 07/02/2020) (cit. on p. 112).

Mélanie-Becquet, Frédérique and Frédéric Landragin (2014). 'Linguistique outillée pour l'étude des chaînes de référence questions méthodologiques et solutions techniques'. In: *Langages* 195

(Sept. 2014), pp. 117–137. URL: https://halshs.archives-ouvertes.fr/halshs-01069462 (visited on 06/07/2017) (cit. on p. 41).

Mikolov, Tomáš (2012). 'Statistical Language Models Based on Neural Networks'. PhD Thesis. Brno, Česká republika: Vysoké Učení Technické v Brně. 133 pp. URL: https://www.fit.vutbr.cz/~imikolov/rnnlm/thesis.pdf (visited on 16/02/2020) (cit. on p. 99).

Mikolov, Tomáš, Kai Chen, Greg Corrado and Jeffrey Dean (2013). 'Efficient Estimation of Word Representations in Vector Space'. In: 1st International Conference on Learning Representations. Ed. by Yoshua Bengio and Yann LeCun. Scottsdale, Arizona, USA, 2nd May 2013. URL: http://arxiv.org/abs/1301.3781 (visited on 28/09/2019) (cit. on pp. 54, 85, 99, 100).

Mikolov, Tomáš, Edouard Grave, Piotr Bojanowski, Christian Puhrsch and Armand Joulin (2018). 'Advances in Pre-Training Distributed Word Representations'. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*. LREC 2018 (Miyazaki, Japan). European Language Resource Association. URL: http://aclweb.org/anthology/L18-1008 (visited on 03/03/2019) (cit. on pp. 99, 100).

Miller, George A. (1995). 'WordNet: a lexical database for English'. In: *Communications of the ACM* 38.11 (1st Nov. 1995), pp. 39–41. DOI: 10.1145/219717.219748. URL: https://doi.org/10.1145/219717.219748 (visited on 29/01/2020) (cit. on p. 51).

Miyamoto, Yasumasa and Kyunghyun Cho (2016). 'Gated Word-Character Recurrent Language Model'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2016. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1992–1997. DOI: 10.18653/v1/D16-1209. URL: https://www.aclweb.org/anthology/D16-1209 (visited on 11/09/2019) (cit. on p. 79).

Moosavi, Nafise Sadat, Leo Born, Massimo Poesio and Michael Strube (2019). 'Using Automatically Extracted Minimum Spans to Disentangle Coreference Evaluation from Boundary Detection'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019 (Firenze, Italia). Association for Computational Linguistics, July 2019, pp. 4168–4178. DOI: 10.18653/v1/P19-1408. URL: https://www.aclweb.org/anthology/P19-1408 (visited on 31/01/2020) (cit. on pp. 20, 21).

Moosavi, Nafise Sadat and Michael Strube (2016). 'Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016. Vol. 1. Berlin, Deutschland: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/P16-1060. URL: http://www.aclweb.org/anthology/P16-1060 (visited on 13/03/2018) (cit. on pp. 20, 26).

Morerio, Pietro, Jacopo Cavazza, Riccardo Volpi, Rene Vidal and Vittorio Murino (2017). *Curriculum Dropout*. 17th Mar. 2017. arXiv: 1703.06229 [cs, stat]. URL: http://arxiv.org/abs/1703.06229 (visited on 06/09/2019) (cit. on p. 81).

MUC Consortium (1995a). 'Appendix C: Named Entity Task Definition (v2.1)'. In: Sixth Message Understanding Conference. Columbia, Maryland: Morgan Kaufmann, 6th Nov. 1995. URL: https://www.aclweb.org/anthology/M95-1024 (visited on 30/11/2019) (cit. on p. 16).

MUC Consortium (1995b). 'Appendix D: Coreference Task Definition (v2.3)'. In: Sixth Message Understanding Conference. Columbia, Maryland: Morgan Kaufmann, 6th Nov. 1995. URL: https://www.aclweb.org/anthology/M95-1025 (visited on 12/11/2019) (cit. on pp. 13, 64).

MUC Consortium (1995c). *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland November 6-8, 1995*. San Francisco, California, USA: Morgan Kaufmann, 6th Nov. 1995. URL: https://www.aclweb.org/anthology/M95-1000 (visited on 12/11/2019) (cit. on pp. 2, 13, 30, 32, 52).

MUC Consortium (1998). *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. San Francisco, California, USA: Morgan Kaufmann Publishers Inc. URL: https://www.aclweb.org/anthology/M98-1000 (visited on 19/12/2019) (cit. on pp. 2, 30, 52).

Munkres, James (1957). 'Algorithms for the Assignment and Transportation Problems'. In: *Journal of the Society for Industrial and Applied Mathematics* 1.5 (Mar. 1957), pp. 32–38 (cit. on p. 24).

Muzerelle, Judith, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau and Iris Eshkol (2013a). 'ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement'. In: *Actes de la 20ème conférence sur le Traitement Automatique des Langues Naturelles*. TALN'2013. Les Sable d'Olonne, France: Association pour le Traitement Automatique des Langues, June 2013, pp. 555–563. URL: https://hal.archives-ouvertes.fr/hal-01016562 (visited on 31/01/2017) (cit. on pp. 3, 35, 36, 65, 111).

Muzerelle, Judith, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol and Jeanne Villaneau (2014). 'ANCOR Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures'. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. LREC 2014 (Reykjavík, Ísland). European Language Resources Association. URL: https://hal.archives-ouvertes.fr/hal-01075679 (visited on 06/07/2017) (cit. on pp. 3, 9, 11–13, 19, 31, 35–37, 41, 45, 65, 67, 71, 111).

Muzerelle, Judith, Emmanuel Schang, Jean-Yves Antoine, Iris Eshkol, Denis Maurel, Aurore Boyer-Pelletier and Damien Nouvel (2013b). 'Annotation en relations anaphoriques d'un corpus de discours oral spontané en français'. In: *Congrès Mondial de Linguistique Française*. CMLF'2012. Lyon, France, July 2013. URL: https://hal.archives-ouvertes.fr/hal-00788164 (visited on 18/12/2019) (cit. on pp. 36, 41).

Nair, Vinod and Geoffrey E. Hinton (2010). 'Rectified Linear Units Improve Restricted Boltzmann Machines'. In: 27th International Conference on International Conference on Machine Learning (Haifa, Israel). ICML'10. Haifa, Israel: Omnipress, June 2010, pp. 807–814. URL: http://dl.acm.org/citation.cfm?id=3104322.3104425 (visited on 30/09/2019) (cit. on p. 88).

Nasr, Alexis, Frédéric Béchet, Jean-François Rey, Benoit Favre and Joseph Le Roux (2011). 'MACAON : An NLP Tool Suite for Processing Word Lattices'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. ACL 2011. United States, pp. 86–91. URL: https://hal.archives-ouvertes.fr/hal-00702442 (visited on 04/03/2019) (cit. on p. 65).

Nedoluzhko, Anna, Michal Novák, Silvie Cinkova, Marie Mikulová and Jiří Mírovský (2016). 'Coreference in Prague Czech-English Dependency Treebank'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC 2016. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association, 23rd–28th May 2016 (cit. on pp. 32, 40).

Ng, Vincent (2008). 'Unsupervised Models for Coreference Resolution'. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2008 (Honolulu, Hawai'i, USA). Association for Computational Linguistics, Oct. 2008, pp. 640–649. URL: https://www.aclweb.org/anthology/D08-1067 (visited on 10/02/2020) (cit. on p. 64).

Ng, Vincent and Claire Cardie (2002). 'Improving Machine Learning Approaches to Coreference Resolution'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL 2002 (Philadelphia, Pennsylvania, USA). Association for Computational Linguistics, July 2002, pp. 104–111. DOI: 10.3115/1073083.1073102. URL: https://www.aclweb.org/anthology/P02-1014 (visited on 29/01/2020) (cit. on pp. 51, 52, 56, 60, 61, 65).

Nicolas, Pascale, Sabine Letellier-Zarshenas, Igor Schadle, Jean-Yves Antoine and Jean Caelen (2002). 'Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations'. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. LREC 2002 (Las Palmas de Gran Canaria, Canarias, España). European Language Resources Association, May 2002. URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/111.pdf (visited on 10/01/2020) (cit. on p. 36).

Nitoń, Bartłomiej, Paweł Morawiecki and Maciej Ogrodniczuk (2018). 'Deep Neural Networks for Coreference Resolution for Polish'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC 2018. Miyazaki, Japan: European Languages Resources Association, May 2018. URL: https://www.aclweb.org/anthology/L18-1060 (visited on 22/09/2019) (cit. on pp. 54, 72).

Niven, Timothy and Hung-Yu Kao (2019). 'Probing Neural Network Comprehension of Natural Language Arguments'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Firenze, Italia: Association for Computational Linguistics, pp. 4658–4664. DOI: 10.18653/v1/P19-1459. URL: https://www.aclweb.org/anthology/P19-1459 (visited on 22/09/2019) (cit. on p. 72).

Nivre, Joakim et al. (2016). 'Universal Dependencies v1: A Multilingual Treebank Collection'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC 2016 (23rd–28th May 2016). Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association, May 2016 (cit. on pp. 45, 103).

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy and James R. Curran (2013). 'Learning multilingual named entity recognition from Wikipedia'. In: *Artificial Intelligence*. Artificial Intelligence, Wikipedia and Semi-Structured Resources 194 (1st Jan. 2013), pp. 151–175. DOI: 10.1016/j.artint.2012.03.006. URL: http://www.sciencedirect.com/science/article/pii/S0004370212000276 (visited on 17/02/2020) (cit. on p. 103).

Ogrodniczuk, Maciej, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary and Magdalena Zawisławska (2016). 'Polish Coreference Corpus'. In: *Human Language Technology. Challenges for Computer Science and Linguistics*. Ed. by Zygmunt Vetulani, Hans Uszkoreit and Marek Kubis. Lecture Notes in Computer Science. Springer International Publishing, pp. 215–226 (cit. on pp. 33, 39, 41, 43).

Ogrodniczuk, Maciej and Vincent Ng, eds. (2016). *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*. San Diego, California: Association for Computational Linguistics, June 2016. DOI: 10.18653/v1/W16-07. URL: https://www.aclweb.org/anthology/W16-0700 (visited on 20/01/2020) (cit. on p. 34).

Ogrodniczuk, Maciej and Vincent Ng, eds. (2017). *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017. DOI: 10.18653/v1/W17-15. URL: https://www.aclweb.org/anthology/W17-1500 (visited on 20/01/2020) (cit. on p. 34).

Ortiz Suárez, Pedro Javier, Yoann Dupont, Benjamin Muller, Laurent Romary and Benoît Sagot (2020). 'Establishing a New State-of-the-Art for Named Entity Recognition'. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*. LREC 2020 (Marseilles, France). European Language Resource Association (cit. on p. 101).

Ortiz Suárez, Pedro Javier, Benoît Sagot and Laurent Romary (2019). 'Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures'. In: *7th Workshop on the Challenges in the Management of Large Corpora*. CMLCè7 (Cardiff, United Kingdom). Ed. by Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen and Caroline Iliadi. Leibniz-Institut für Deutsche Sprache, July 2019. URL: https://hal.inria.fr/hal-02148693 (visited on 31/01/2020) (cit. on pp. 54, 101).

Pāṇini (–0349). *Aṣṭādhyāyī* (cit. on p. 106).

Paszke, Adam et al. (2017). 'Automatic differentiation in PyTorch'. In: NeurIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques. Long Beach, CA, 9th Dec. 2017. URL: https://openreview.net/forum?id=BJJsrmfCZ (visited on 19/08/2019) (cit. on p. 80).

Pennington, Jeffrey, Richard Socher and Christopher Manning (2014). 'Glove: Global Vectors for Word Representation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2014 (Doha, Qatar). Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162 (visited on 16/02/2020) (cit. on p. 100).

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer (2018). 'Deep Contextualized Word Representations'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL: HLT 2018 (New Orleans, Louisiana). Vol. 1. New Orleans, Louisiana, USA: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: http://aclweb.org/anthology/N18-1202 (visited on 16/02/2019) (cit. on pp. 72, 85, 101).

Plank, Barbara, Anders Søgaard and Yoav Goldberg (2016). 'Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016. Vol. 2. Berlin, Germany: Association for Computational Linguistics, pp. 412–418. DOI: 10.18653/v1/P16-2067. URL: http://aclweb.org/anthology/P16-2067 (visited on 03/09/2019) (cit. on p. 72).

Poesio, Massimo (2004). 'The MATE/GNOME Proposals for Anaphoric Annotation, Revisited'. In: *Proceedings of the 5th SIGDIAL Workshop*. Boston, Massachusetts, USA, pp. 154–162. URL: http://cswww.essex.ac.uk/staff/poesio/publications/SIGDIAL04.pdf (visited on 06/07/2017) (cit. on p. 40).

Poesio, Massimo (2016). 'Linguistic and Cognitive Evidence About Anaphora'. In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuck-

ardt and Yannick Versley. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 23–54. DOI: 10.1007/978-3-662-47909-4_2. URL: https://doi.org/10.1007/978-3-662-47909-4_2 (visited on 24/11/2019) (cit. on p. 19).

Poesio, Massimo, Mijail Alexandrov-Kabadjov, Renata Vieira, Rodrigo Goulart and Olga Uryupina (2005). 'Does discourse-new detection help definite description resolution'. In: *Proceedings of the 6th International Workshop on Computational Semantics*. IWCS 2005 (Tilburg, Nederland), pp. 236–246 (cit. on pp. 52, 78).

Poesio, Massimo and Ron Artstein (2008). 'Anaphoric Annotation in the ARRAU Corpus'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC 2008. Marrakech, Morocco: European Language Resources Association. URL: http://ict.usc.edu/pubs/Anaphoric%20Annotation%20in%20the%20ARRAU%20Corpus.pdf (cit. on pp. 31, 40, 64).

Poesio, Massimo, Florence Bruneseaux and Laurent Romary (1999). 'The MATE meta-scheme for coreference in dialogues in multiple languages'. In: *ACL'99 Workshop Towards Standards and Tools for Discourse Tagging*. College Parc, United States, June 1999, pp. 65–74. URL: https://hal.inria.fr/inria-00525171 (visited on 06/07/2017) (cit. on p. 40).

Poesio, Massimo, Sameer Pradhan, Marta Recasens, Kepa Rodriguez and Yannick Versley (2016a). 'Annotated Corpora and Annotation Tools'. In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt and Yannick Versley. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 97–163. DOI: 10.1007/978-3-662-47909-4_2. URL: https://doi.org/10.1007/978-3-662-47909-4_2 (visited on 24/11/2019) (cit. on pp. 34, 35).

Poesio, Massimo, Ron Stuckardt and Yannick Versley, eds. (2016b). *Anaphora Resolution: Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer-Verlag Berlin Heidelberg. URL: http://www.springer.com/fr/book/9783662479087 (visited on 06/12/2016) (cit. on p. 22).

Poesio, Massimo, Ron Stuckardt, Yannick Versley and Renata Vieira (2016c). 'Early Approaches to Anaphora Resolution: Theoretically Inspired and Heuristic-Based'. In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt and Yannick Versley. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 55–94. DOI: 10.1007/978-3-662-47909-4_2. URL: https://doi.org/10.1007/978-3-662-47909-4_2 (visited on 24/11/2019) (cit. on p. 17).

Poesio, Massimo et al. (2018). 'Anaphora Resolution with the ARRAU Corpus'. In: *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*. CRAC 2018 (New Orleans, Louisiana). New Orleans, Louisiana, USA: Association for Computational Linguistics, 6th June 2018, pp. 11–22. DOI: 10.18653/v1/W18-0702. URL: http://aclweb.org/anthology/W18-0702 (visited on 01/03/2019) (cit. on pp. 64, 89).

Popescu-Belis, Andrei (1999). 'Modélisation multi-agents des échanges langagiers : application au problème de la référence et à son évaluation'. thesis. Paris 11, 1st Jan. 1999. URL: http://www.theses.fr/1999PA112200 (visited on 21/01/2020) (cit. on pp. 35, 36).

Popescu-Belis, Andrei, Loïs Rigouste, Susanne Salmon-Alt and Laurent Romary (2004). 'Online Evaluation of Coreference Resolution.' In: *4th International Conference on Language Resources and Evaluation*. LREC 2004. Lisboa, Portugal: European Language Resources Association.

URL: https://halshs.archives-ouvertes.fr/halshs-00005023/document (visited on 01/03/2019) (cit. on p. 20).

Pradhan, Sameer, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel (2007a). 'OntoNotes: A Unified Relational Semantic Representation'. In: *Proceedings of the First IEEE International Conference on Semantic Computing*. ICSC 2007 (Irvine, California, USA). IEEE Computer Society, pp. 517–526. DOI: 10.1109/ICSC.2007.67. URL: http://dx.doi.org/10.1109/ICSC.2007.67 (visited on 06/07/2017) (cit. on pp. 5, 64).

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng and Michael Strube (2014). 'Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL 2014. Vol. 2. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 30–35. URL: http://www.aclweb.org/anthology/P14-2006 (cit. on pp. 24, 26).

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina and Yuchen Zhang (2012). 'CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes'. In: *Proceedings of the Joint EMNLP-CoNLL conference*. EMNLP-CoNLL 2012 (□□□, □□□□ (Jeju Island, Republic of Korea)). Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, pp. 1–40. URL: http://aclweb.org/anthology/W12-4501 (visited on 28/02/2019) (cit. on pp. 30, 33, 64).

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel and Nianwen Xue (2011). 'CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes'. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. CoNLL 2011 (Portland, Oregon, USA). CONLL Shared Task '11. Association for Computational Linguistics, pp. 1–27. URL: http://dl.acm.org/citation.cfm?id=2132936.2132937 (visited on 01/10/2017) (cit. on pp. 30, 33, 64).

Pradhan, Sameer, Lance Ramshaw, Ralph Weischedel, Jessica MacBride and Linnea Micciulla (2007b). 'Unrestricted Coreference: Identifying Entities and Events in OntoNotes'. In: *Proceedings of the First IEEE International Conference on Semantic Computing*. ICSC 2007 (Irvine, California, USA). Minneapolis, Minnesota, USA, Sept. 2007, pp. 446–453. DOI: 10.1109/ICSC.2007.93 (cit. on pp. 19, 34, 39).

Prince, Ellen Friedman (1981). 'Toward a taxonomy of given - new information'. In: *Radical pragmatics*. Ed. by Peter Cole. New York: Academic Press, pp. 223–255 (cit. on pp. 3, 8, 9, 12, 40).

Quignard, Matthieu, Serge Heiden, Frédéric Landragin and Matthieu Decorde (2018). 'Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes'. In: *Proceedings of the Fourteenth International Conference on the Statistical Analysis of Textual Data*. JADT 2018. UniversItalia, 11th June 2018, pp. 610–615. URL: https://hal.archives-ouvertes.fr/hal-01814858 (visited on 04/02/2020) (cit. on p. 47).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever (2019). 'Language models are unsupervised multitask learners' (cit. on p. 81).

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky and Christopher D. Manning (2010). 'A Multi-pass Sieve for Coreference Resolution'. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2010 (Cambridge, Massachusetts, USA). Association for Com-

putational Linguistics, pp. 492–501. URL: http://dl.acm.org/citation.cfm?id=1870658.1870706 (visited on 31/01/2017) (cit. on pp. 63, 96, 97).

Rahman, Altaf and Vincent Ng (2009). 'Supervised models for coreference resolution'. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2009 (Singapore). Vol. 2. Association for Computational Linguistics, p. 968. DOI: 10.3115/1699571.1699639. URL: http://portal.acm.org/citation.cfm?doid=1699571.1699639 (visited on 29/09/2019) (cit. on p. 85).

Rahman, Altaf and Vincent Ng (2011). 'Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution'. In: *Journal of Artificial Intelligence Research* 40 (25th Feb. 2011), pp. 469–521. DOI: 10.1613/jair.3120. URL: https://www.jair.org/index.php/jair/article/view/10694 (visited on 30/01/2020) (cit. on p. 63).

Rand, William M. (1971). 'Objective Criteria for the Evaluation of Clustering Methods'. In: *Journal of the American Statistical Association* 66.336, pp. 846–850. DOI: 10.2307/2284239. JSTOR: 2284239 (cit. on p. 25).

Rastier, François (2002). 'Enjeux épistémologiques de la linguistique de corpus'. In: *Deuxièmes journées de la linguistique de corpus*. Ed. by Williams G. Lorient, France: Presses Universitaires de Rennes, Sept. 2002, pp. 31–46. URL: https://hal.archives-ouvertes.fr/hal-00171532 (visited on 19/12/2019) (cit. on p. 30).

Recasens, Marta (2010). 'Coreference: Theory, Annotation, Resolution and Evaluation.' Universitat de Barcelona. URL: http://stel.ub.edu/cba2010/phd/phd.pdf (visited on 18/10/2016) (cit. on pp. 14, 20, 21, 78).

Recasens, Marta, Marie-Catherine de Marneffe and Christopher Potts (2013). 'The Life and Death of Discourse Entities: Identifying Singleton Mentions'. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL: HLT 2013 (Atlanta, Georgia, USA). Association for Computational Linguistics, June 2013, pp. 627–633. URL: https://www.aclweb.org/anthology/N13-1071 (visited on 29/01/2020) (cit. on p. 53).

Recasens, Marta and Eduard Hovy (2009). 'A Deeper Look into Features for Coreference Resolution'. In: *Anaphora Processing and Applications*. Ed. by Sobha Lalitha Devi, António Branco and Ruslan Mitkov. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 29–42. DOI: 10.1007/978-3-642-04975-0_3 (cit. on p. 52).

Recasens, Marta and Eduard Hovy (2010). 'Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL 2010 (Uppsala, Sverige, 11th–16th July 2010). Uppsala, Sverige: Association for Computational Linguistics, pp. 1423–1432. URL: http://www.aclweb.org/anthology/P10-1144 (visited on 31/01/2017) (cit. on pp. 17, 63).

Recasens, Marta and Eduard Hovy (2011). 'BLANC: Implementing the Rand Index for Coreference Evaluation'. In: *Natural Language Engineering* 17.4 (Oct. 2011), pp. 485–510. DOI: 10.1017/S135132491000029X. URL: http://dx.doi.org/10.1017/S135132491000029X (cit. on pp. 20, 23–25).

Recasens, Marta, Eduard Hovy and M. Antònia Martí (2011a). 'Identity, non-identity, and near-identity: Addressing the complexity of coreference'. In: *Lingua* 121.6 (1st May 2011), pp. 1138–

1152. DOI: 10.1016/j.lingua.2011.02.004. URL: http://www.sciencedirect.com/science/article/pii/S0024384111000325 (visited on 05/08/2019) (cit. on pp. 10, 19, 42).

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio and Yannick Versley (2010). 'SemEval-2010 task 1: Coreference resolution in multiple languages'. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval 2010 (Uppsala, Sverige). Association for Computational Linguistics, pp. 1–8 (cit. on pp. 26, 33, 64, 112).

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio and Yannick Versley (2011b). *SemEval-2010 task 1: Coreference resolution in multiple languages*. Dataset release LDC2011T01. Philadelphia, Pennsylvania, USA: Linguistic Data Consortium. URL: https://catalog.ldc.upenn.edu/LDC2011T01 (visited on 20/01/2020) (cit. on p. 35).

Reddi, Sashank J., Satyen Kale and Sanjiv Kumar (2018). 'On the Convergence of Adam and Beyond'. In: Sixth International Conference on Learning Representations. Vancouver, Canada, 30th Apr. 2018. URL: https://openreview.net/forum?id=ryQu7f-RZ (visited on 19/08/2019) (cit. on p. 80).

Rodríguez, Kepa Joseba, Francesca Delogu, Yannick Versley, Egon W. Stemle and Massimo Poesio (2010–0021). 'Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. LREC 2010. Ed. by Khalid Choukri, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias and Nicoletta Calzolari. Valletta, Malta: European Language Resources Association, 19th May 2010–21 (cit. on p. 40).

Roesiger, Ina, Maximilian Köper, Kim Anh Nguyen and Sabine Schulte im Walde (2018). 'Integrating Predictions from Neural-Network Relation Classifiers into Coreference and Bridging Resolution'. In: *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 44–49. DOI: 10.18653/v1/W18-0705. URL: https://www.aclweb.org/anthology/W18-0705 (visited on 18/01/2020) (cit. on p. 77).

Rogers, Anna (2019). *How the Transformers broke NLP leaderboards*. 30th June 2019. URL: https://hackingsemantics.xyz/2019/leaderboards/ (visited on 22/09/2019) (cit. on p. 73).

Romary, Laurent (2017). *stdfSpec : A proposal for a stand-off element for the TEI Guidelines*. 7th July 2017. URL: https://github.com/laurentromary/stdfSpec (cit. on p. 44).

Rosset, Sophie, Cyril Grouin and Pierre Zweigenbaum (2011). *Entités nommées structurées : guide d'annotation Quaero*. Notes LIMSI 2011-04. Orsay, France: Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, 8th Sept. 2011, p. 86. URL: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf (visited on 21/11/2019) (cit. on p. 16).

Roth, Dan and Wen-tau Yih (2007). 'Global Inference for Entity and Relation Identification via a Linear Programming Formulation'. In: *Introduction to Statistical Relational Learning*. Ed. by Lise Getoor and Ben Taskar. The MIT Press. DOI: 10.7551/mitpress/7432.003.0022. URL: https://direct.mit.edu/books/book/3811/chapter/125093/global-inference-for-entity-and-relation (visited on 09/02/2020) (cit. on p. 61).

Rubenstein, Herbert and John B. Goodenough (1965). 'Contextual correlates of synonymy'. In: *Communications of the ACM* 8.10 (1st Oct. 1965), pp. 627–633. DOI: 10.1145/365628.365657. URL: https://doi.org/10.1145/365628.365657 (visited on 16/02/2020) (cit. on p. 99).

Ruder, Sebastian (2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. 15th June 2017. arXiv: 1706.05098 [cs, stat]. URL: http://arxiv.org/abs/1706.05098 (visited on 26/03/2019) (cit. on p. 86).

Ruder, Sebastian (2019). 'Neural Transfer Learning for Natural Language Processing'. PhD Thesis. Galway, Éire: National University of Ireland. URL: http://ruder.io/thesis/ (cit. on p. 85).

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (1986). 'Learning representations by back-propagating errors'. In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0 (visited on 17/09/2019) (cit. on p. 80).

Russell, Bertrand (1919). 'The philosophy of logical atomism'. In: *The Monist* 29.1, pp. 32–63. JSTOR: 27900724 (cit. on p. 9).

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi (2019). *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*. 24th July 2019. arXiv: 1907.10641 [cs]. URL: http://arxiv.org/abs/1907.10641 (visited on 30/07/2019) (cit. on p. 18).

Salimans, Tim and Diederik P. Kingma (2016). *Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks*. 25th Feb. 2016. arXiv: 1602.07868 [cs]. URL: http://arxiv.org/abs/1602.07868 (visited on 10/09/2019) (cit. on p. 82).

Salmon-Alt, Susanne (2002). 'Le projet ANANAS : Annotation anaphorique pour l'analyse sémantique de corpus'. In: *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*. TALN 2002 (24th–27th June 2002). Nancy, France. URL: http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=14215371 (visited on 21/01/2020) (cit. on pp. 35, 36).

Salmon-Alt, Susanne, Eckhard Bick, Laurent Romary and Jean-Marie Pierrel (2004). 'La FReeBank : vers une base libre de corpus annotés'. In: Traitement Automatique des Langues Naturelles - TALN'04. Apr. 2004, 10 p. URL: https://hal.inria.fr/inria-00100194 (visited on 21/01/2020) (cit. on pp. 35, 36).

Sanh, Victor, Thomas Wolf and Sebastian Ruder (2018). *A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks*. 14th Nov. 2018. arXiv: 1811.06031 [cs]. URL: http://arxiv.org/abs/1811.06031 (visited on 28/02/2019) (cit. on pp. 86, 112).

Santurkar, Shibani, Dimitris Tsipras, Andrew Ilyas and Aleksander Madry (2018). *How Does Batch Normalization Help Optimization?* 29th May 2018. arXiv: 1805.11604 [cs, stat]. URL: http://arxiv.org/abs/1805.11604 (visited on 10/09/2019) (cit. on p. 83).

Schmid, Hellmut (1994). 'Probabilistic Part-of-Speech Tagging Using Decision Trees'. In: *Proceedings of the International Conference on New Methods in Language Processing*. International Conference on New Methods in Language Processing. Manchester, UK. URL: https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (cit. on p. 40).

Schuster, Mike and Kuldip K. Paliwal (1997). 'Bidirectional recurrent neural networks'. In: *IEEE Transactions on Signal Processing* 45.11 (Nov. 1997), pp. 2673–2681. DOI: 10.1109/78.650093 (cit. on pp. 6, 73).

Sekine, Satoshi and Chikashi Nobata (2004). 'Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy'. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. LREC 2004 (Lisbon, Portugal). European Language Resources Association, May 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/65.pdf (visited on 24/02/2020) (cit. on p. 16).

Sennrich, Rico, Barry Haddow and Alexandra Birch (2016). 'Neural Machine Translation of Rare Words with Subword Units'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016 (Berlin, Deutschland). Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://www.aclweb.org/anthology/P16-1162 (visited on 16/02/2020) (cit. on p. 102).

Sha, Fei and Fernando Pereira (2003). 'Shallow Parsing with Conditional Random Fields'. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. HLT-NAACL 2003 (Edmonton, Canada), pp. 213–220. URL: https://www.aclweb.org/anthology/N03-1028 (visited on 17/02/2020) (cit. on p. 105).

Shen, Dinghan et al. (2018). 'Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL 2018. Vol. 1. Melbourne, Australia: Association for Computational Linguistics, pp. 440–450. DOI: 10.18653/v1/P18-1041. URL: http://aclweb.org/anthology/P18-1041 (visited on 22/09/2019) (cit. on p. 72).

Shimodaira, Hidetoshi (2000). 'Improving predictive inference under covariate shift by weighting the log-likelihood function'. In: *Journal of Statistical Planning and Inference* 90.2 (1st Oct. 2000), pp. 227–244. DOI: 10.1016/S0378-3758(00)00115-4. URL: http://www.sciencedirect.com/science/article/pii/S0378375800001154 (visited on 10/09/2019) (cit. on p. 83).

Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper and Bryan Catanzaro (2019). *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. 17th Sept. 2019. arXiv: 1909.08053 [cs]. URL: http://arxiv.org/abs/1909.08053 (visited on 27/09/2019) (cit. on p. 81).

Simpson, George Gaylord (1947). 'Holarctic Mammalian Faunas and Continental Relationships during the Cenozoic'. In: *Geological Society of America Bulletin* 58, p. 613. DOI: 10.1130/0016-7606(1947)58[613:HMFACR]2.0.CO;2. URL: http://adsabs.harvard.edu/abs/1947GSAB...58..613S (visited on 16/02/2020) (cit. on p. 98).

Smith, Aaron, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao and Sara Stymne (2018). '82 Treebanks, 34 Models: Universal Dependency Parsing with Multi-Treebank Models'. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL 2018 (Bruxelles, Belgique). Association for Computational Linguistics, Oct. 2018, pp. 113–123. DOI: 10.18653/v1/K18-2011. URL: https://www.aclweb.org/anthology/K18-2011 (visited on 07/02/2020) (cit. on p. 112).

Smyth, Ron (1994). 'Grammatical determinants of ambiguous pronoun resolution'. In: *Journal of Psycholinguistic Research* 23.3 (1st Sept. 1994), pp. 197–229. DOI: 10.1007/BF02139085. URL: https://doi.org/10.1007/BF02139085 (visited on 05/02/2020) (cit. on p. 56).

Søgaard, Anders and Yoav Goldberg (2016). 'Deep multi-task learning with low level tasks supervised at lower layers'. In: 54th Annual Meeting of the Association for Computational

Linguistics. ACL 2016. Vol. 2. Berlin, Deutschland: Association for Computational Linguistics, Aug. 2016, pp. 231–235. DOI: 10.18653/v1/P16-2038. URL: http://www.aclweb.org/anthology/P16-2038 (visited on 26/03/2019) (cit. on p. 86).

Song, Zhiyi et al. (2015). 'From Light to Rich ERE: Annotation of Entities, Relations, and Events'. In: 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Denver, Colorado: Association for Computational Linguistics, pp. 89–98. DOI: 10.3115/v1/W15-0812. URL: http://aclweb.org/anthology/W15-0812 (visited on 24/11/2019) (cit. on p. 18).

Soon, Wee Meng, Hwee Tou Ng and Chung Yong Lim (2001). 'A Machine Learning Approach to Coreference Resolution of Noun Phrases'. In: *Computational Linguistics* 27.4, pp. 521–544. DOI: 10.1162/089120101753342653. URL: https://www.aclweb.org/anthology/J01-4004/ (visited on 14/08/2019) (cit. on pp. 51–53, 56, 59, 60, 65, 98).

Soraluze, Ander, Olatz Arregi, Xabier Arregi, Klara Ceberio and Arantza Díaz de Ilarraza (2012). 'Mention detection: First steps in the development of a Basque coreference resolution system'. In: *Proceedings of KONVENS 2012*. Ed. by Jeremy Jancsary. ÖGAI, Sept. 2012, pp. 128–136. URL: http://www.oegai.at/konvens2012/proceedings/18_soraluze12o/ (visited on 30/11/2016) (cit. on pp. 33, 40, 65, 76, 105).

Soraluze, Ander, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza (2019). 'EUSKOR: End-to-end coreference resolution system for Basque'. In: *PLOS ONE* 14.9 (12th Sept. 2019), e0221801. DOI: 10.1371/journal.pone.0221801. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221801 (visited on 17/09/2019) (cit. on p. 53).

Soraluze, Ander, Olatz Arregi, Xabier Arregi and Arantza Díaz De Ilarraza (2017). 'Improving mention detection for Basque based on a deep error analysis'. In: *Natural Language Engineering* 23.3 (May 2017), pp. 351–384. DOI: 10.1017/S1351324916000206. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/improving-mention-detection-for-basque-based-on-a-deep-error-analysis/89C8669493FC0942CC32013BE57D8C0C (visited on 17/09/2019) (cit. on pp. 65, 76, 89).

Sørensen, Thorvald Julius (1948). *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. København: E. Munksgaard (cit. on p. 24).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov (2014). 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. In: *Journal of Machine Learning Research* 15, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html (visited on 06/09/2019) (cit. on pp. 81, 82).

Stern, Mitchell, Jacob Andreas and Dan Klein (2017). 'A Minimal Span-Based Neural Constituency Parser'. In: 55th Annual Meeting of the Association for Computational Linguistics. Vol. 1. Vancouver, Canada: Association for Computational Linguistics, pp. 818–827. DOI: 10.18653/v1/P17-1076. URL: http://aclweb.org/anthology/P17-1076 (visited on 28/02/2019) (cit. on p. 74).

Stevenson, Rosemary J., Alexander W. R. Nelson and Keith Stenning (1995). 'The Role of Parallelism in Strategies of Pronoun Comprehension'. In: *Language and Speech* 38.4 (1st Oct. 1995), pp. 393–418. DOI: 10.1177/002383099503800404. URL: https://doi.org/10.1177/002383099503800404 (visited on 05/02/2020) (cit. on p. 56).

Stoyanov, Veselin, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler and David Hysom (2010). 'Coreference Resolution with Reconcile'. In: *Proceedings of the ACL 2010 Conference*. ACL 2010 (Uppsala, Sverige). Vol. Short papers. Association for Computational Linguistics, July 2010, pp. 156–161. URL: https://www.aclweb.org/anthology/P10-2029 (visited on 30/01/2020) (cit. on p. 65).

Stoyanov, Veselin and Jason Eisner (2012). 'Easy-first Coreference Resolution'. In: URL: https://core.ac.uk/display/22102734 (visited on 10/02/2020) (cit. on p. 64).

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie and Ellen Riloff (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 ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. ACL-AFNLP 2009 (Suntec, Singapore). Association for Computational Linguistics, Aug. 2009, pp. 656–664. URL: https://www.aclweb.org/anthology/P09-1074 (visited on 30/11/2019) (cit. on p. 64).

Straka, Milan and Jana Straková (2017). 'Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe'. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL 2017. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 88–99. DOI: 10.18653/v1/K17-3009. URL: https://www.aclweb.org/anthology/K17-3009 (visited on 23/01/2020) (cit. on p. 45).

Straka, Milan and Jana Straková (2019). 'ÚFAL MRPipe at MRP 2019: UDPipe Goes Semantic in the Meaning Representation Parsing Shared Task'. In: *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*. CoNLL 2019 (Hong Kong). Association for Computational Linguistics, Nov. 2019, pp. 127–137. DOI: 10.18653/v1/K19-2012. URL: https://www.aclweb.org/anthology/K19-2012 (visited on 17/02/2020) (cit. on p. 103).

Straková, Jana, Milan Straka and Jan Hajic (2019). 'Neural Architectures for Nested NER through Linearization'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Firenze, Italia: Association for Computational Linguistics, July 2019, pp. 5326–5331. URL: https://www.aclweb.org/anthology/P19-1527 (visited on 06/08/2019) (cit. on p. 16).

Sturt, Patrick and Vincenzo Lombardo (2005). 'Processing Coordinated Structures: Incrementality and Connectedness'. In: *Cognitive Science* 29.2, pp. 291–305. DOI: 10.1207/s15516709cog0000_8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_8 (visited on 08/02/2020) (cit. on p. 58).

Sukhbaatar, Sainbayar, Arthur Szlam, Jason Weston and Rob Fergus (2015). 'End-To-End Memory Networks'. In: *Advances in Neural Information Processing Systems*. NeurIPS 2015 (Montréal, Canada). Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett. Vol. 28. Curran Associates, Inc., pp. 2440–2448. URL: http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf (visited on 06/02/2020) (cit. on p. 57).

Sutskever, Ilya, Oriol Vinyals and Quoc V Le (2014). 'Sequence to Sequence Learning with Neural Networks'. In: *Advances in Neural Information Processing Systems*. 28th conference on Neural Information Processing. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger. Vol. 27. Montréal, Canada: Curran Associates, Inc., pp. 3104–3112. URL:

http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (visited on 21/09/2019) (cit. on p. 74).

Swayamdipta, Swabha, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer and Noah A. Smith (2018). 'Syntactic Scaffolds for Semantic Structures'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Association for Computational Linguistics, Oct. 2018, pp. 3772–3782. DOI: 10.18653/v1/D18-1412. URL: https://www.aclweb.org/anthology/D18-1412 (visited on 19/02/2020) (cit. on p. 112).

Szymkiewicz, Dezydery (1934). *Une Contribution statistique à la géographie floristique*. Polskie Towarzystwo Botaniczne. 17 pp. Google Books: AK13oAEACAAJ (cit. on p. 98).

Taulé, Mariona, M. Antònia Martí and Marta Recasens (2008). 'AnCora: Multilevel Annotated Corpora for Catalan and Spanish'. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morroco: European Language Resources Association, May 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf (visited on 06/07/2017) (cit. on pp. 26, 32, 40).

TEI consortium (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.0*. Guidelines. TEI consortium, 13th Feb. 2020. URL: http://www.tei-c.org/Guidelines/P5/ (visited on 09/10/2017) (cit. on pp. 5, 43).

Tellier, Isabelle, Yoann Dupont and Arnaud Courmet (2012). 'Un segmenteur-étiqueteur et un chunker pour le français'. In: *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*. TALN 2012 (Grenoble, France). Vol. 5. Association pour le Traitement Automatique des Langues, June 2012, pp. 7–8. URL: https://www.aclweb.org/anthology/F12-5004 (visited on 17/02/2020) (cit. on p. 106).

Tellier, Isabelle, Yoann Dupont, Iris Eshkol and Ilaine Wang (2013). 'Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort Is Necessary?' In: *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning*. IDEAL 2013, Special Session on Text Data Learning (□□, □□ (Héféi, China)). Ed. by Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li and Xin Yao. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 226–233. DOI: 10.1007/978-3-642-41278-3_28. URL: https://hal.archives-ouvertes.fr/hal-01174605 (cit. on p. 106).

Tellier, Isabelle, Iris Eshkol-Taravella, Yoann Dupont and Ilaine Wang (2014). 'Peut-on bien chunker avec de mauvaises étiquettes POS ?' In: *Actes de la 21ème conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2014. TALN2014. Marseille, France: Association pour le Traitement Automatique des Langues, July 2014, pp. 125–136. URL: https://hal.archives-ouvertes.fr/hal-01024274 (visited on 13/04/2017) (cit. on pp. 68, 106).

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister and Kathrin Beck (2006). *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)*. Technical report. Tubingen, Germany: Universitat Tubingen, (cit. on pp. 39, 40).

Trouilleux, François (2001). 'Referential links identification and automatic interpretation of pronominal expressions in French texts'. Theses. Université Blaise-Pascal, Clermont-Ferrand, Dec. 2001. URL: https://hal.archives-ouvertes.fr/tel-01152394 (visited on 12/04/2017) (cit. on pp. 36, 65, 67).

Turian, Joseph, Lev-Arie Ratinov and Yoshua Bengio (2010). 'Word Representations: A Simple and General Method for Semi-Supervised Learning'. In: *Proceedings of the 48th Annual Meeting*

of the Association for Computational Linguistics. ACL 2010 (Uppsala, Sverige). Association for Computational Linguistics, July 2010, pp. 384–394. URL: https://www.aclweb.org/anthology/P10-1040 (visited on 16/02/2020) (cit. on p. 99).

Tutin, Agnès, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot and Georges Antoniadis (2000). 'Annotating a large corpus with anaphoric links'. In: *Proceedings of the 3rd International Conference on Discourse Anaphora and Anaphora Resolution*. DAARC 2000. Lancaster, UK. URL: https://hal.archives-ouvertes.fr/hal-00373327 (visited on 04/09/2017) (cit. on pp. 35, 36).

Uryupina, Olga (2003). 'High-precision Identification of Discourse New and Unique Noun Phrases'. In: *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*. ACL 2003 (Sapporo, Japan). Association for Computational Linguistics, July 2003, pp. 80–86. DOI: 10.3115/1075178.1075189. URL: https://www.aclweb.org/anthology/P03-2012 (visited on 29/01/2020) (cit. on pp. 52, 78).

Uryupina, Olga (2004). 'Evaluating Name-Matching for Coreference Resolution'. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. LREC 2004 (Lisbon, Portugal). European Language Resources Association, May 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/376.pdf (visited on 29/01/2020) (cit. on p. 52).

Uryupina, Olga (2006). 'Coreference Resolution with and without Linguistic Knowledge'. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. LREC 2006 (Genova, Italia). European Language Resources Association, May 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/726_pdf.pdf (visited on 29/01/2020) (cit. on p. 52).

Uryupina, Olga (2007). 'Knowledge acquisition for coreference resolution'. PhD Thesis. Saarbrücken, Deutschland: Universitäat des Saarlandes. 280 pp. DOI: http://dx.doi.org/10.22028/D291-23499. URL: https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/23555 (visited on 29/01/2020) (cit. on p. 52).

Uryupina, Olga (2010). 'Corry: A System for Coreference Resolution'. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval 2010 (Uppsala, Sverige). Association for Computational Linguistics, pp. 100–103. URL: http://dl.acm.org/citation.cfm?id=1859664.1859684 (visited on 01/03/2019) (cit. on pp. 61, 65, 89).

Uryupina, Olga, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez and Massimo Poesio (2019). 'Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus'. In: *Natural Language Engineering* (7th May 2019), pp. 1–34. DOI: 10.1017/S1351324919000056. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/annotating-a-broad-range-of-anaphoric-phenomena-in-a-variety-of-genres-the-arrau-corpus/17E7FA2CB2E36C213E2649479593B6B0 (visited on 22/12/2019) (cit. on pp. 31, 32, 40, 64).

Uryupina, Olga, Mijail Kabadjov and Massimo Poesio (2016). 'Detecting Non-reference and Non-anaphoricity'. In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt and Yannick Versley. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 23–54. DOI: 10.1007/978-3-662-47909-4_2. URL: https://doi.org/10.1007/978-3-662-47909-4_2 (visited on 24/11/2019) (cit. on p. 20).

Uryupina, Olga and Alessandro Moschitti (2013). 'Multilingual Mention Detection for Coreference Resolution'. In: *Proceedings of the sixth International Joint Conference on Natural Language Processing*. IJCNLP 2013 (Nagoya, Japan). Asian Federation of Natural Language Processing, Oct. 2013, pp. 100–108. URL: https://www.aclweb.org/anthology/I13-1012 (visited on 30/11/2019) (cit. on pp. 20, 32, 65).

Van Deemter, Kees and Rodger Kibble (2000). 'On Coreferring: Coreference in MUC and Related Annotation Schemes'. In: *Computational Linguistics* 26.4, pp. 629–637. URL: https://www.aclweb.org/anthology/J00-4005/ (cit. on p. 32).

Van Rijsbergen, Cornelis Joost (1979). *Information Retrieval*. Butterworth-Heinemann, 1st Jan. 1979. URL: http://dl.acm.org/citation.cfm?id=539927 (visited on 30/11/2019) (cit. on p. 20).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. Long Beach, California: Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (visited on 16/02/2019) (cit. on pp. 73, 83, 101).

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*. ACL-08: HLT (Columbus, Ohio, USA, 15th–20th June 2008). Association for Computational Linguistics, pp. 9–12. URL: http://dl.acm.org/citation.cfm?id=1564144.1564147 (cit. on p. 51).

Vieira, Renata and Massimo Poesio (2000). 'An Empirically-based System for Processing Definite Descriptions'. In: *Computational Linguistics* 26.4, pp. 539–593. DOI: 10.1162/089120100750105948. URL: https://www.aclweb.org/anthology/J00-4003 (visited on 29/01/2020) (cit. on pp. 52, 78).

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly and Lynette Hirschman (1995). 'A Model-Theoretic Coreference Scoring Scheme'. In: *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, pp. 45–52. URL: http://www.aclweb.org/anthology/M95-1005 (visited on 02/03/2018) (cit. on pp. 14, 21, 23).

Villemonte De La Clergerie, Éric, Benoît Sagot and Djamé Seddah (2017). 'The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy'. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL 2017 (Vancouver, Canada). Aug. 2017, pp. 243–252. DOI: 10.18653/v1/K17-3026. URL: https://hal.inria.fr/hal-01584168 (cit. on pp. 45, 107).

Vinyals, Oriol, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever and Geoffrey Hinton (2015). 'Grammar As a Foreign Language'. In: *Proceedings of the 28th International Conference on Neural Information Processing* (Montreal, Canada). Vol. 2. NIPS'15. Cambridge, MA, USA: MIT Press, pp. 2773–2781. URL: http://dl.acm.org/citation.cfm?id=2969442.2969550 (visited on 16/02/2019) (cit. on p. 74).

Wan, Li, Matthew Zeiler, Sixin Zhang, Yann Le Cun and Rob Fergus (2013). 'Regularization of Neural Networks using DropConnect'. In: *International Conference on Machine Learning*. International Conference on Machine Learning. 13th Feb. 2013, pp. 1058–1066. URL: http://proceedings.mlr.press/v28/wan13.html (visited on 06/09/2019) (cit. on p. 81).

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel R. Bowman (2019). *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. 12th July 2019. arXiv: 1905.00537 [cs]. URL: http://arxiv.org/abs/1905.00537 (visited on 30/11/2019) (cit. on p. 18).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel Bowman (2018). 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. BlackboxNLP 2018. Bruxelles, Belgique: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: http://aclweb.org/anthology/W18-5446 (visited on 30/11/2019) (cit. on p. 18).

Wang, Bailin, Wei Lu, Yu Wang and Hongxia Jin (2018). 'A Neural Transition-based Model for Nested Mention Recognition'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018 (Brussels, Belgium). Bruxelles, Belgique: Association for Computational Linguistics, pp. 1011–1017. URL: http://aclweb.org/anthology/D18-1124 (visited on 28/02/2019) (cit. on pp. 65, 76, 94, 112).

Wang, Wenhui and Baobao Chang (2016). 'Graph-based Dependency Parsing with Bidirectional LSTM'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL 2016. Vol. 1. Berlin, Deutschland: Association for Computational Linguistics, pp. 2306–2315. DOI: 10.18653/v1/P16-1218. URL: http://aclweb.org/anthology/P16-1218 (visited on 21/09/2019) (cit. on p. 74).

Warde-Farley, David, Ian J. Goodfellow, Aaron Courville and Yoshua Bengio (2013). *An empirical analysis of dropout in piecewise linear networks*. 20th Dec. 2013. arXiv: 1312.6197 [cs, stat]. URL: http://arxiv.org/abs/1312.6197 (visited on 06/09/2019) (cit. on p. 81).

Webber, Bonnie Lynn (1978). 'Description Formation and Discourse Model Synthesis'. In: Theoretical Issues in Natural Language Processing-2. Urbana, Illinois, 25th July 1978. URL: https://www.aclweb.org/anthology/T78-1006 (cit. on pp. 8, 9).

Webber, Bonnie Lynn (1991). 'Structure and ostension in the interpretation of discourse deixis'. In: *Language and Cognitive Processes* 6.2 (1st Apr. 1991), pp. 107–135. DOI: 10.1080/01690969108406940. URL: https://doi.org/10.1080/01690969108406940 (visited on 24/11/2019) (cit. on p. 19).

Weber, Florence and Stéphane Beaud (2003). *Guide de l'enquête de terrain : produire et analyser des données ethnographiques*. Paris: Éditions la Découverte. 356 pp. (cit. on p. 37).

Webster, Kellie, Marta Recasens, Vera Axelrod and Jason Baldridge (2018). 'Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns'. In: *Transactions of the Association for Computational Linguistics* 6 (1st Dec. 2018), pp. 605–617. DOI: 10.1162/tacl_a_00240. arXiv: 1810.05201. URL: https://doi.org/10.1162/tacl_a_00240 (visited on 28/07/2020) (cit. on pp. 18, 31, 73).

Weischedel, Ralph et al. (2013). *OntoNotes release 5.0 with OntoNotes DB Tool v0.999 beta*. Dataset release LDC2013T19. Philadelphia: Linguistic Data Consortium, p. 53. URL: https://catalog.

ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf (visited on 04/08/2019) (cit. on pp. 33, 34).

Weiss, Gail, Yoav Goldberg and Eran Yahav (2018). 'On the Practical Computational Power of Finite Precision RNNs for Language Recognition'. In: 56th Annual Meeting of the Association for Computational Linguistics. Vol. 2. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 740–745. URL: https://www.aclweb.org/anthology/P18-2117 (visited on 23/07/2019) (cit. on pp. 71, 73, 96).

Weng, Lilian (2019). *Are Deep Neural Networks Dramatically Overfitted?* 14th Mar. 2019. URL: https://lilianweng.github.io/2019/03/14/are-deep-neural-networks-dramatically-overfitted.html (visited on 24/09/2019) (cit. on p. 81).

Weston, Jason, Sumit Chopra and Antoine Bordes (2015). 'Memory Networks'. In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR 2015 (San Diego, California, USA). 29th Nov. 2015. arXiv: 1410.3916. URL: http://arxiv.org/abs/1410.3916 (visited on 06/02/2020) (cit. on p. 57).

Widlöcher, Antoine (2008). 'Analyse macro-sémantique des structures rhétoriques du discours : cadre théorique et modèle opératoire'. thesis. Caen, 1st Jan. 2008. URL: http://www.theses.fr/2008CAEN2042 (visited on 22/01/2020) (cit. on p. 41).

Widlöcher, Antoine and Yann Mathet (2012). 'The Glozz Platform: A Corpus Annotation and Mining Tool'. In: 2012 ACM Symposium on Document Engineering. DocEng '12. New York, NY, USA: ACM, pp. 171–180. DOI: 10.1145/2361354.2361394. URL: http://doi.acm.org/10.1145/2361354.2361394 (visited on 06/07/2017) (cit. on pp. 41, 42).

Winograd, Terry (1972). 'Understanding natural language'. In: *Cognitive Psychology* 3.1 (1st Jan. 1972), pp. 1–191. DOI: 10.1016/0010-0285(72)90002-3. URL: http://www.sciencedirect.com/science/article/pii/0010028572900023 (visited on 20/02/2020) (cit. on p. 2).

Wiseman, Sam, Alexander M. Rush and Stuart M. Shieber (2016). 'Learning Global Features for Coreference Resolution'. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL: HLT 2016 (San Diego, California). Association for Computational Linguistics, pp. 994–1004. DOI: 10.18653/v1/N16-1114. URL: http://aclweb.org/anthology/N16-1114 (visited on 16/02/2019) (cit. on pp. 53, 57, 63, 84, 94).

Wiseman, Sam, Alexander M. Rush, Stuart M. Shieber and Jason Weston (2015). 'Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. ACL-IJCNLP 2015 (□□□, □□ (Beijing, China), 26th–31st July 2015). Vol. 1. Association for Computational Linguistics, pp. 1416–1426. URL: http://aclweb.org/anthology/P/P15/P15-1137.pdf (cit. on pp. 53, 61, 63, 78, 84, 109).

Wong, Catherine, Neil Houlsby, Yifeng Lu and Andrea Gesmundo (2018). 'Transfer Learning with Neural AutoML'. In: *Advances in Neural Information Processing Systems 31*. 2018 conference on Neural Information Processing Systems. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. Montréal, Canada: Curran Associates, Inc., 2nd Dec. 2018, pp. 8356–8365. URL: http://papers.nips.cc/paper/8056-transfer-learning-with-neural-automl.pdf (visited on 06/10/2019) (cit. on p. 93).

Yang, Xiaofeng, Jian Su, Guodong Zhou and Chew Lim Tan (2004). 'An NP-cluster Based Approach to Coreference Resolution'. In: *Proceedings of the 20th International Conference on Computational Linguistics*. ACL 2004 (Barcelona, España). COLING '04. Association for Computational Linguistics. DOI: 10.3115/1220355.1220388. URL: https://doi.org/10.3115/1220355.1220388 (visited on 30/01/2017) (cit. on pp. 57, 63).

Yang, Xiaofeng, Guodong Zhou, Jian Su and Chew Lim Tan (2003). 'Coreference Resolution Using Competition Learning Approach'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. ACL 2003 (Sapporo, Japan). Association for Computational Linguistics, July 2003, pp. 176–183. DOI: 10.3115/1075096.1075119. URL: https://www.aclweb.org/anthology/P03-1023 (visited on 09/02/2020) (cit. on p. 59).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht and Oriol Vinyals (2016). *Understanding deep learning requires rethinking generalization*. 10th Nov. 2016. arXiv: 1611.03530 [cs]. URL: http://arxiv.org/abs/1611.03530 (visited on 24/09/2019) (cit. on p. 81).

Zhang, Chiyuan, Samy Bengio and Yoram Singer (2019). 'Are All Layers Created Equal?' In: *Proceedings of the workshop on Identifying and Understanding Deep Learning Phenomena*. 36th International Conference on Machine Learning. Long Beach, California, 5th Feb. 2019. arXiv: 1902.01996. URL: http://arxiv.org/abs/1902.01996 (visited on 06/10/2019) (cit. on p. 93).

Zhang, Rui, Cicero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang and Dragomir Radev (2018). 'Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering'. In: 56th Annual Meeting of the Association for Computational Linguistics. ACL 2018. Vol. 2. Melbourne, Australia, July 2018, pp. 102–107. URL: https://aclweb.org/anthology/papers/P/P18/P18-2017/ (visited on 01/04/2019) (cit. on p. 77).

Zidouni, Azeddine, Sophie Rosset and Hervé Glotin (2010). 'Efficient combined approach for named entity recognition in spoken language'. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association*. InterSpeech 2010 (Makuhari, Chiba, Japan). International Speech Communication Association (cit. on p. 106).

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Oxford, England: Addison-Wesley Press (cit. on p. 70).

Zoph, Barret and Quoc V. Le (2017). 'Neural Architecture Search with Reinforcement Learning'. In: 5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 26th Apr. 2017. URL: https://openreview.net/forum?id=r1Ue8Hcxg (visited on 06/10/2019) (cit. on p. 93).

List of Figures

3.1 Syntactic analysis (subtree) for "*au moment où je me suis marié en juillet soixante-sept*" 47

5.1 DeCOFre inference mode operating process 70
5.2 Intrinsic words representations in DeCOFre 71
5.3 Contextual words representations in DeCOFre 72
5.4 Boundaries and context embeddings in DeCOFre 74
5.5 Soft-head self-attention in DeCOFre 74

List of Tables

2.1 Definition of the BLANC metric 26

3.1 French corpora with anaphora annotations prior 2002 (from Salmon-Alt 2002) 36
3.2 ANCOR subcorpora dimensions (from Muzerelle et al. 2014) 37

5.1 Embedding dimensions in DeCOFre baseline 87
5.2 Feedfoward networks parameters in DeCOFre baseline 88
5.3 Recurrent layers parameters in DeCOFre baseline 89
5.4 Mention detection evaluation (ANCOR test) 89
5.5 Coreference resolution evaluation (ANCOR test) 90
5.6 Influence of training setting on mention detection 90
5.7 Influence of the training setting on coreference resolution 91
5.8 Influence of the learning rate schedule on mention detection 92
5.9 Influence of weight decay on mention detection performances 92
5.10 Influence of undersampling rate on mention detection perfomances 93

6.1 Influence of the length feature on mention detection (ANCOR dev) 97
6.2 Influence of the length feature on coreference resolution (ANCOR dev) 97
6.3 Influence of string matching features on coreference resolution (ANCOR dev) 99
6.4 Influence of word embeddings on mention detection (ANCOR dev) 102
6.5 Influence of word embeddings on coreference resolution (ANCOR dev) 102
6.6 Influence of shallow linguistic knowledge on mention detection (ANCOR dev) 104
6.7 Influence of shallow linguistic knowledge on coreference resolution (ANCOR dev) 104
6.8 Influence of structural linguistic knowledge on mention detection (ANCOR dev) 108
6.9 Influence of structural linguistic knowledge on coreference resolution (ANCOR dev) 108

Licence

This document is available under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0)
(https://creativecommons.org/licenses/by/4.0/deed.en)

Copyright © 2020, Loïc Grobol <loic.grobol@gmail.com>



This work is part of the 'Investissements d'Avenir' overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL), and is also supported by the ANR DEMOCRAT (Describing and Modelling Reference Chains: Tools for Corpus Annotation and Automatic Processing) project ANR-15-CE38-0008 and the ANR Profiterole (PRocessing Old French Instrumented TExts for the Representation Of Language Evolution) project ANR-16-CE38-0010.

Reconnaissance automatique de chaînes de coréférences en français parlé

Une *chaîne de coréférences* est l'ensemble des expressions linguistiques — ou *mentions* — qui font référence à une même entité ou un même objet du discours. La tâche de reconnaissance des chaînes de coréférences consiste à détecter l'ensemble des mentions d'un document et à le partitionner en chaînes de coréférences. Ces chaînes jouent un rôle central dans la cohérence des documents et des interactions et leur identification est un enjeu important pour de nombreuses autres tâches en traitement automatique du langage, comme l'extraction d'informations ou la traduction automatique. Des systèmes automatiques de reconnaissance de chaînes de coréférence existent pour plusieurs langues, mais aucun pour le français ni pour une langue parlée.

Nous nous proposons dans cette thèse de combler ce manque par un système de reconnaissance automatique de chaînes de coréférences pour le français parlé. À cette fin, nous proposons un système utilisant des réseaux de neurones artificiels et ne nécessitant pas de resources externes. Ce système est viable malgré le manque d'outils de prétraitements adaptés au français parlé et obtient des performances comparable à l'état de l'art. Nous proposons également des voies d'amélioration de ce système, en y introduisant des connaissances issues de ressources et d'outils conçus pour le français écrit. Enfin, nous proposons un nouveau format de représentation pour l'annotation des chaînes de coréférences dans des corpus de langues écrites et parlées et en nous en donnons un exemple en proposant under nouvelle version d'ANCOR — le premier corpus de français annoté en coréférence.

Mots-clés : anaphore, coréférence, réseaux de neurones artificiels, apprentissage artificiel, ressources annotées, corpus, formats d'annotation, traitement automatique du langage naturel

Coreference resolution for spoken French

A *coreference chain* is the set of linguistic expressions — or *mentions* — that refer to the same entity or discourse object in a given document. *Coreference resolution* consists in detecting all the mentions in a document and partitioning their set in coreference chains. Coreference chains play a central role in the consistency of documents and interactions, and their identification has applications to many other fields in natural language processing that rely on understanding of language, such as information extraction, question answering or machine translation. Natural language processing systems that perform this task exist for many languages, but none for French — which suffered until recently from a lack of suitable annotated resources — and none for spoken language.

In this thesis, we aim to fill this lack by designing a coreference resolution system for spoken French. To this end, we propose a knowledge-poor system based on an end-to-end neural network architecture, which obviates the need for the preprocessing pipelines common in existing systems, while maintaining performances comparable to the state-of-the art. We then propose extensions on that baseline, by augmenting our system with external knowledge obtained from resources and preprocessing tools designed for written French. Finally, we propose a new standard representation for coreference annotation in corpora of written and spoken languages, and demonstrate its use in a new version of ANCOR, the first coreference corpus of spoken French.

Keywords: anaphora, coreference, artificial neural networks, machine learning, annotated resources, corpus, annotations representation, natural language processing

École Doctorale 622 — Sciences du Langage,
Université Sorbonne Nouvelle, maison de la recherche
4, rue des Irlandais, 75005 Paris