Kingdom of Morocco
Mohamed First University
Faculty of Sciences Oujda

المملكة المغربية
جامعة محمد الأول
كلية العلوم وجدة

مركز دراسات الدكتوراه علوم وتقنيات

Centre for Doctoral Studies Science and Technology

Thesis

Order n°: **492/18**

Submitted and presented by:

Imad Zeroual

In accordance with the requirements for the degree

DOCTOR OF SCIENCE

Doctoral Program: Mathematics and Computer Science

Specialty: Computer Science

Building Arabic Corpora: Concepts, Methodologies, Tools, and Experiments

The thesis is publicly defended on 24/11/2018 in front of a Dissertation Committee composed of:

Prof. El Mostafa Daoudi	Faculty of Sciences Oujda	Chair
Prof. Violetta Cavalli-Sforza	Al Akhawayn University, Ifrane	Reviewer
Prof. Karim Bouzoubaa	Mohammadia School of Engineers, Rabat	Reviewer
Prof. Tim Buckwalter	University of Maryland, USA	Reviewer
Prof. Azzeddine Mazroui	Faculty of Sciences Oujda	Examiner
Prof. Abdelouafi Meziane	Faculty of Sciences Oujda	Examiner
Prof. Abdelhak Lakhouaja	Faculty of Sciences Oujda	Thesis Supervisor

Acknowledgments

In the name of Allah, the most Merciful and Beneficent

First and Foremost, praise is to Allah, the Almighty, the greatest of all, on whom ultimately, we depend for sustenance and guidance. I would like to thank Allah for giving me health, knowledge, ability, and opportunity to undertake this research study, and to persevere and complete it satisfactorily.

This thesis would not have seen the light without the enormous and endless support, guidance, and encouragement of my supervisor, Prof. *Abdelhak Lakhouaja*, for which my mere expression of thanks likewise does not suffice.

Special thanks to Prof. *Azzeddine Mazroui*, Prof. *Abdelouafi Meziane*, Prof. *Rachid Belahbib*, and Prof. *Abderrahim Boudlal*, who acted as extra supervisors. They provided expert guidance on Arabic morphological and syntactic theory and how best to approach related problems using machine learning. I benefited and learned immensely from them how to be a real researcher, and what I needed to make original contributions to new areas of research.

I owe my deepest appreciation to the reviewers, Prof. *Violetta Cavalli-Sforza*, Prof. *Tim Buckwalter,* and Prof. *Karim Bouzoubaa*, for their high-quality reviews, which provided insightful and constructive comments. I am very grateful to them for their belief in the direction and quality of my work which provide the motivation I needed to see my thesis through to completion. I would like to thank Prof. *El Mostafa Daoudi*, who is the Chair of the Program Committee members. His help and discussions are the key points for a successful thesis defence.

I would like to thank my fellow PhD students and my friends at Laboratory of Computer Science, especially NLP group members for the great seminars and training courses that we used to enjoy every year. They each helped make my time during this PhD program more fun and interesting. Moreover, having *Anoual El KAh* as a colleague, a co-author, a beloved wife, and a kind mother of my new born daughter is a great privilege and joy, and I would like to thank her for believing in me and my goals.

Finally, I dedicate this thesis to my family who have always supported me in my studies and life. Without your love, care and patience, I would not have achieved this. The special dedication of this thesis is to the most beloved Mum. Thank you for your patience, care and everything you have done to keep our family gathered in peace and happiness. Thank you for giving us the love we need to survive in this life. I always love you Mum.

Abstract

The term corpus comes from Latin and means “body”. According to corpus linguists, a corpus can be defined as a collection of machine-readable authentic texts, including transcripts of spoken data. The focus of corpora builders is essentially divided into three areas: corpus compilation, data processing, and corpus annotation. Each one of these tasks requires specialists, takes time, and costs money. The further task is to infer information from corpora to provide empirical evidence for linguistic theories or to turn the data into products or services. Corpora are essential resources for computational linguistics and Natural Language Processing (NLP) fields. Expressly, corpora include empirical data that enable linguists and grammarians to form objective rather than subjective statements. Further, many NLP applications are moving from rule-based systems and knowledge-based methods to data-driven approaches.

The prime motivation for carrying out the research in this thesis comes from the limited research on Arabic corpus linguistics and the lack of available resources, standards, and efficient tools that can cope with the perspectives of Arabic NLP. Furthermore, most Arabic corpora builders have often proposed corpora and tools that comply with their suitable objectives without considering the standardization and the international aspects. Therefore, another purpose of this thesis is to provide an overview of central criteria and methodology of building corpora and to give a better understanding of Arabic corpus linguistics.

To widen the scope of this thesis, it was necessary to carry out some tasks:

1) We conducted a survey that covers 100 well-known and influential corpora to know how relevant corpora have been built, yet, what and how long it takes to complete the procedure. The survey presents a summarisation of data sources and different compilation methods used in relation to corpus characteristics like size and time consumed during the compilation process.

2) Basically, there is a lack of appropriate tools that can deal effectively with the richness of morphology and syntax of both Classic and Modern Standard Arabic (MSA). Thus, we developed our own tools and adapted others namely stemmer, lemmatizer, and part-of-speech tagger. In doing so, we study and investigate the state-of-the-art of available tools, then, we propose standard concepts and tagset considering the Arabic language features. Furthermore, we carefully collect Arabic linguistics resources to create the required dictionaries to enhance the performance of developed and adapted tools. Finally, comparative and usability tests are performed.

3) In order to enrich our work, we built three different types of corpora: Classic Arabic (i.e., Al-Mus’haf), MSA (i.e., OSIAN), and multilingual (i.e., MulTed). Detailed information about the building procedures and the characteristics of the constructed corpora are presented. Furthermore, they are compared to similar corpora, stressing their significant contribution to the literature. Finally, these corpora will publicly release to push forward the state-of-the-art in Arabic NLP and corpus linguistics.

ملخص

أصل مصطلح كوربوس CORPUS (الذّخيرة فيما بعد) لاتيني، ويعني "الجسد"، وهو تجميع منهجي مصمّم ومنظّم للنّصوص. يمكن تعريف الذّخيرة وفقا للّسانيين باعتبارها مجموعة من النّصوص المكتوبة، أو نسخا لبيانات منطوقة بحيث تكون قابلة للاستدعاء والقراءة من قبل الآلة. ينقسم تركيز مطوّري الذّخائر إلى ثلاثة مجالات أساسية: تجميع النّصوص، ومعالجتها، ثمّ توسيمها. وكل واحدة من هذه المهام تتطلّب اختصاصيين، وتستغرق وقتا، وتكلّف جهدا ومالا. ويتبع تجميع النصوص، استنباط المعلومات من الذّخيرة لتوفير أدلّة تجريبية مباشرة للنّظريات اللّغوية، كما يمكن توظيف هذه المعلومات في تطوير منتجات أخرى. تُعتبر المعالجة الآلية للّغات الطّبيعية واللّغويات الحاسوبية من بين المجالات التي تعتمد كثيرا على الذّخائر، ويرجع ذلك أساسًا إلى أنّ العديد من تطبيقات المعالجة الآلية للّغات الطّبيعية انتقلت من اعتمادها الكلّي على الخوارزميات المستندة على القواعد اللّغوية إلى استعمال أساليب قائمة على التّدريب الإحصائي الذي يعتمد على البيانات المتوفرة في الذّخائر. كما تُعتبر الذّخائر مصادر لبيانات تجريبية تمكّن اللّغويين والنّحاة من صياغة دلائل موضوعية أكثر منها ذاتية.

يرجع الدّافع الرّئيسي لإنجاز أطروحة علمية في هذا المجال إلى محدودية الأبحاث المُنجزة وقلتها، إضافة إلى قلّة الموارد والأدوات المتاحة التي يمكنها التّعامل بكفاءة مع حاجيات المعالجة الآلية للّغة العربية وتطلّعاتها. لذلك، فإن هدفنا الأساسي يكمن في تقديم نظرة شمولية حول المفاهيم الأساسية والأساليب المتّبعة لبناء وتطوير ذخائر باللّغة العربية.

لتوسيع نطاق الأطروحة، كان من الضروري إنجاز مايلي:

1. إجراء دراسة استقصائية تغطي 100 ذخيرة مشهورة لمعرفة تفاصيل وحيثيات بناء تلك الذخائر حيث قدّمنا ملخصًا لمصادر البيانات وطرق التجميع المختلفة المستخدمة مقترنة بخصائص مثل الحجم والوقت المستهلك خلال عملية التجميع.
2. نظرا لقلة الأدوات القادرة على التعامل بكفاءة مع خصائص الصرف النحو العربي قمنا بتصميم الأدوات وإعداد الموارد اللّازمة لبناء هذه الذخائر بما يتناسب مع خصائص اللّغة العربية سواء في النّصوص القديمة أو الحديثة. ولتحقيق هذه الغاية، قمنا بدراسة ومعاينة أحدث ما تمّ إنجازه من تقنيات ووسائل لبناء الذّخائر اللّغوية، ثم طوّرنا بدورنا برامج، واقترحنا معايير وأساليب تعالج وتتخطّى سلبيات ومشاكل ما هو متوفر حاليا لأجل الارتقاء والتّقدم في مجال اللّغويات الحاسوبية للّغة العربية.
3. تجميع ثلاث ذخائر مختلفة منها من يحتوي على النّص القرآني (Al-Mus'haf)، وأخرى على النّصوص الحديثة (OSIAN) والأخيرة متعددة اللغات (MulTed) ثم قمنا بتقديم معلومات تفصيلية حول عملية التجميع بالإضافة إلى عرض خصائص هذه الذخائر ومقارنتها مع نظيرتها من الذخائر. أخيرا، سوف سوف يتم إصدار هذه الذخائر مجانا للمساهمة في تقدم مجال المعالجة الحاسوبية للغة العربية.

Résumé

Le terme corpus vient du latin et signifie « corps ». Selon les linguistes de corpus, un corpus peut être défini comme une collection de textes authentiques lisibles à la machine comprenant des transcriptions de données parlées. L'objectif des constructeurs de corpus est essentiellement divisé en trois domaines : la compilation de corpus, le traitement de données et l'annotation de corpus. Chacune de ces tâches nécessite un personnel qualifié, prend du temps et coûte de l'argent. Il s'ensuit que l'inférence des informations provenant des corpus fournit des preuves expérimentales directes des théories linguistiques ou transforme les données en produits. Le Traitement Automatiques des langues Naturelles (TALN) et la linguistique computationnelle sont fortement parmi les domaines influencés par les corpus. Fondamentalement, au lieu d'utiliser les règles et les méthodes basées sur les connaissances, la plupart des applications TALN utilisent les approches basées sur l'apprentissage. De plus, les corpus sont considérés comme des données empiriques qui permettent aux linguistes et grammairiens de former des énoncés objectifs plutôt que subjectifs.

La principale motivation pour effectuer la recherche dans cette thèse vient des recherches limitées sur la linguistique des corpus arabes et le manque de ressources disponibles, des standards et d'outils efficaces qui peuvent faire face aux perspectives du traitement automatique de la langue arabe. Par conséquent, nos objectifs sont de fournir une vue globale, de l'ensemble des concepts et des méthodes centraux de construction de corpus et de donner une meilleure compréhension de la linguistique du corpus arabe.

Pour élargir la portée de cette thèse, il était nécessaire de réaliser certaines tâches :

4) Nous avons mené une enquête portant sur 100 corpus bien connus et influents pour savoir comment les corpus pertinents ont été construits. Cette enquête présente un résumé des sources de données et des différentes méthodes de compilation utilisées en relation avec les caractéristiques du corpus telles que la taille et le temps passé au cours du processus de compilation.

5) En fait, il y a un manque d'outils appropriés qui peuvent traiter efficacement la richesse de la morphologie et la syntaxe de l'Arabe classique et standard moderne. Ainsi, nous avons proposé des normes et des méthodes qui tiennent compte des caractéristiques de la langue Arabe. Ensuite, nous avons collecté soigneusement les textes en arabe, créé les dictionnaires et les ressources linguistiques nécessaires pour améliorer la performance des outils développés. Ces outils ont été adaptés pour le stemming, la lemmatisation et la partie du discours qui marquent les données compilées. Enfin, des tests comparatifs et d'utilisabilité sont effectués.

6) Afin d'enrichir notre travail, nous avons construit trois types de corpus : pour l'arabe classique (Al-Mus'haf), pour l'Arabe standard moderne (OSIAN) et la dernière est multilingue (MulTed). Des informations détaillées sur les procédures de construction et les caractéristiques des corpus construits sont présentées. En outre, ils sont comparés à des corpus similaires, soulignant leur importante contribution à la littérature. Enfin, ces corpus seront rendus publics pour faire avancer l'état de l'art en le Traitement Automatique de la Langue Arabe (TALA) et en linguistique de corpus.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
ALC	Arabic Learner Corpus
BAMA	Buckwalter Arabic Morphological Analyzer
BNC	British National Corpus
BoE	Bank of English
CA	Classical Arabic
CCA	Corpus of Contemporary Arabic
CLIR	Cross-Lingual Information Retrieval
CTS	Corpus-based Translation Studies
DOI	Digital Object Identifier
EAGLES	Expert Advisory Group on Language Engineering Standards
HMMs	Hidden Markov Models
IAMA	Improved Arabic Morphology Analyser
ICE	International Corpus of English
IR	Information Retrieval
KACST	King Abdulaziz City for Science and Technology
KSUCCA	King Saud University Corpus of Classical Arabic
LOB	Lancaster, Oslo, and Bergen
MSA	Modern Standard Arabic
NLP	Natural Language Processing
OSAC	Open Source Arabic Corpora
OSIAN	Open Source International Arabic News
PATB	Penn Arabic TreeBank
PoS	Part of Speech
QALB	Qatar Arabic Language Bank
SGML	Standard Generalized Markup Language
SVM	Support Vector Machine
WARC	Web ARChive format
WSJ	Wall Street Journal
XML	eXtensible Markup Language

CHAPTER 1: Introduction

1. Introduction

This chapter aims to provide an introduction dedicated to the topic of this thesis. Further, it discusses the impact of corpus-based studies on various fields research. Then, our objectives and methodologies are presented. Next, we stress the motivation, highlighting the reasons behind the choice of the Arabic language. Finally, the structure of this thesis is presented.

2. This Thesis

Corpora are the core of any scientific field that is based on extensive human linguistic data. Among these fields, we mention computational linguistics, Natural Language Processing (NLP), and education. The title "Building Arabic corpora: Concepts, Methodologies, Tools, and Experiments" reflects, to some extent, the overall objectives covered by this thesis. What is important here is that the language addressed is Arabic, one of the six United Nation official languages since 1974. The limits of existing Arabic corpora may be explained by the lack of available tools that efficiently deal with the complex morphology and the linguistic specificities of Arabic. Further, the proposed corpora and developed tools typically comply with their builders' objectives without considering standardization and international aspects. In addition, major progress is found in using data driven approaches that paves the way for highly effective NLP applications. This improvement is due mainly to large and high-quality corpora. Unfortunately, the lack of such resources, either for training or evaluating tasks, affects the performance of Arabic NLP applications.

The general concern of this thesis is to provide guidelines, standards, techniques, and tools to help corpora builders to design and build reliable and reference corpora that push forwards the advance of Arabic corpus linguistics. Further, we aim to respond to today's challenges, and to cope with the demands, expectations, and perspectives of different NLP research groups.

In this work, the long-established Arabic grammar and its linguistics features have been taken into consideration. Moreover, the evaluation tests involved vowelized and non-vowelized texts from a wide range of formats, domains, and genres, of both Modern Standard and Classical Arabic. Besides, the results are reviewed by Arabic linguistics experts.

3. Impact of Corpora

This section moves on to discuss the impact of corpora on various fields such as linguistic, lexicography, language teaching, and NLP.

Corpora are, in essence, a source of evidence for linguistic description and argumentation. Grammarians have always needed sources of evidence as a basis to illustrate aspects of language such as the nature, the structure, and the functions of language. For instance, Watson (2002) has noted that the Arabic language was codified primarily in the Quran; also, it was based on the language of the western Hijazi tribe of Quraysh, with some interference from pre-Islamic eastern dialects and poetic koiné which was an "inter-Arabic" or a "standard spoken Arabic" that provides the basis of intercommunication between Arabs of different countries (Ferguson 1959). These sources have been used as examples to illustrate Arabic grammatical features or construction. Furthermore, Leech (1992a) claims that using corpora allows observation of

language in use, which leads to theories rather than vice versa. Besides, Halliday et al. (2014) stress that corpus-based analysis is an important source of insight into the nature of language, and is specifically geared to investigate frequencies in corpora to establish probabilities in the grammatical system in order to understand language variation and grammatical change across registers. It is worth mentioning, that before the invention of the computer, building a word frequency list was not a trivial task. For instance, to publish the first known word frequency list in 1897, it needed the help of over five thousand assistants over a period of years to process a German corpus of 11 million tokens (Bongers 1947; Kennedy 2014).

In addition, corpora are exceedingly useful in lexicography (Teubert 2015). They are used to describe reliably the lexicon as well as the grammar of a language. They support many aspects of dictionary creation such as developing a headword list, writing individual entries and identifying their syntactic behaviour, discovering words senses, providing examples and translations (Kilgarriff 2013). For instance, historical dictionaries are among the leading corpus-based dictionaries, which aim to encompass the entire lexicon of a language throughout its history by listing every word and its meanings from its first appearance in written texts to the present (e.g., Historical Dictionary of Arabic[1]). Besides, major revision of relevant dictionaries is systematically based on corpora (Milfull 2009), such as the "Dictionary of the Older Scottish Tongue", the "Middle English Dictionary", the "Dictionary of Old English", and the "Oxford English Dictionary".

Alongside the linguistic description and lexicography, corpora significantly affect a wide range of research activities that have a pedagogical purpose. For instance, word frequency lists are intended to gather statistical information on the use of words and letters of a language. Several researchers emphasized the potential relevance of corpus linguistics for language learning and teaching in all its forms and uses (Boulton and Landure 2016; Bertels 2017). In terms of pedagogy, they believe that corpus linguistics should be considered for use in education to reduce the time that would be necessary to learn a language. Further, they report that corpora are successfully used as a reference resource by both advanced learners as well as learners with lower levels of proficiency or needing language for specific purposes.

As known, the progress in most empirical and statistical approaches used in NLP is driven by available data. Thus, large and high-quality corpora become very valuable resources and many research groups have been concerned with the use of corpora in a variety of NLP applications (Armstrong et al. 2013). For instance, beneficial effects of corpora have been observed in several NLP tasks such as word sense disambiguation (Lefever and Hoste 2013), summarization (Li et al. 2013), syntactic annotation (Xing et al. 2016), and named entity recognition (Nothman et al. 2013). According to our view, machine translation is the NLP application that stands to benefit the most from corpus linguistics (cf. (Hu and others 2016)). In fact, the overlapping between corpus linguistics and descriptive translation studies have contributed to the birth and rise of the corpus-based translation studies (CTS). CTS have become a major research methodology that applies statistical analysis of words or phrases in parallel multilingual corpora to obtain probabilities of translations. Moreover, *Hu* (2016) has explained how corpora can be used in teaching translation, primarily on the establishment of the corpus-based mode of translation teaching and the use of corpora in compiling translation textbooks.

[1] https://www.dohadictionary.org/EN/pages/default.aspx

4. Objectives and Methodology

In general, the aim of this thesis is to outline the main stages in corpus building, from corpus design and compilation to corpus processing and analysis. Basically, the purpose for which a corpus is compiled influences its design, size, and nature. Further, specialists usually build corpora that comply with their objectives; yet, some corpora have been designed for general purposes. Therefore, to be able to compile or use corpora successfully, some factors must necessarily be taken into consideration; otherwise, the results may be different from the expected. Knowing the motivations and aims of corpus building, and understanding the nature of corpus, are among those factors.

The main stages of the corpus building procedure and its related methods that we covered in this thesis are corpus design, compilation, data processing, and corpus annotation. Several issues are associated with each stage. For instance, corpus designers focus on design criteria in order to create a well-defined corpus that meets the standards. To systematically develop a corpus, the latter must be balanced considering the genre of the included texts, the topics and domains covered, and the size of the corpus, among other criteria. Furthermore, some scientific groups work on developing methods and tools for data compilation and processing. Finally, one of the main aims being addressed by corpora builders is to develop new forms of annotation and improve the accuracy of automatic annotation. To sum up, our objectives regarding each corpus building stage are as follows:

1. Corpus design criteria: developing different types of corpora.
2. Selection of sources: given greater attention to sampling and representativeness.
3. Corpus compilation: preparing texts compiled in appropriate machine-readable forms.
4. Text processing/handling: developing effective tool that can deal with the richness of morphology of both Classic and Modern Standard Arabic.
5. Corpus annotation: Proposing a standard tagset for Arabic language, yet, adapting and enhancing relevant part-of-speech tagging methods to annotate Arabic texts.

5. Motivation and Context

In the last decade, the amount of available data grew significantly and many projects on building large corpora have been launched. Unfortunately, not all languages have benefited equally from this growth. The Arabic is an example of such languages. It is expanding in the world in an area extending from the Arabian/Persian Gulf in the East to the Atlantic Ocean in the West. According to UNESCO, the Arabic language is used by more than 422 million persons (Bokova 2012) around 29 countries. Further, the presence of the Arabic language on the web grew around 7,247.3% in the last seventeen years (2000–2017) scoring the highest growth of the ten top online languages, yet, Arabic is the fourth most used language on the web[2]. Around 1.6 billion Muslims worldwide use Arabic to perform their daily prayers in which 80% of them are not Arabic native speakers (Yassein and Wahsheh 2016); also, due to cultural and commercial perspectives, teaching Arabic as a foreign language has become a global educational enterprise (Sakho 2012). Moreover, Arabic is an old Semitic language, i.e., the standardization of its lexicon and grammar are deeply rooted and well established a long time ago in history.

[2] http://www.internetworldstats.com/stats7.htm

Typically, a resource-poor language refers to the language that lacks "the basic resources that are fundamental to computational linguistics" and has a few and small corpora (Zamin et al. 2012). Despite this proud heritage, lexical richness, and online users' growth, Arabic is relatively an under-resourced language compared to other languages with less or similar population size (e.g., French and German). What's more, until 2011, Rabiee (2011) reported that not a single modern standard Arabic tagged corpus was freely or publicly available. Several factors may explain the limits of Arabic corpora building projects such as the inefficient tools developed or adapted to deal with Arabic linguistic features which differ from Indo-European languages. Further, most Arabic corpora builders have often proposed corpora and tools that comply with their suitable objectives without considering the standardization and the international aspects. Moreover, the majority of tagsets used are derived from English, which is a drawback for a morphologically complex language such as Arabic. It is well known that the creation of valuable corpora is expensive. Thus, another factor is the absence of funding and investment for the development of free large and well-defined Arabic corpora.

All these reasons captured our attention, as many researchers and academia, to this field to bridge the gap between Arabic and other resource-rich languages (e.g., English) that have a vast number of resources and tools. Some considerable efforts have been (and are) performed; however, the absence of standards, free resources, and efficient tools that take into consideration the Arabic features are still the main challenges faced by Arabic corpora builders. The chief purpose of this work is to lead those builders into the right directions to refine their own corpora rather than to blindly follow procedures applied to other languages.

6. Thesis Structure

The previous sections have introduced corpora building procedure and their impact on many fields. It also discussed the motivations and objectives. The remainder of this thesis is arranged in four main chapters focusing on the basic aspects of corpus building with experiences on the Arabic language.

In Chapter 2, we present background information and a literature review. In addition to a brief introduction, Section 2 provides the early history of corpora and the roots of what some would consider the first developed corpus, also, the first corpus to be given that name. Section 3 illustrates the design criteria and the corpus characteristics to build a well-defined corpus that meets the standards. Further, corpus typology is described focusing on relevant types of corpora and their primary uses. Section 4 provides detailed information on corpus compilation task. To make this section equally rich in both theoretical and practical aspects, it is supported by results of a survey that covers 100 well-known and influential corpora. Expressly, the survey presents a summarisation of data sources and different compilation methods used in relation to corpus characteristics like size and time consumed during the compilation process. Section 5 addresses the Arabic language features, followed by an overview of the major progress achieved in building Arabic corpora. Finally, we conclude this chapter in Section 6.

In Chapter 3, text handling processes are described, each of which has been developed to deal with morphological structure. For morphology, stemming and lemmatization are two essential morphological analyses widely used in NLP and information retrieval. Basically, this Chapter is divided on two main sections; the first one addresses the stemming task and the second one for the lemmatization. These both techniques are first introduced, illustrated, evaluated, and their uses are investigated throughout various subsections.

Chapter 4 is dedicated to corpus annotation, primarily Part-of-Speech (PoS) tagging. The latter demands certain requirements to be fulfilled, which are described in Section 2. Further, many PoS tagging methods are introduced in Section 3. Section 4 is dedicated to tagsets, and a standard PoS tagset for Arabic is proposed and evaluated. In Section 5, relevant statistical language-independent PoS taggers are presented and adapted to Arabic. Then, a comparative study is performed to conclude a combination method that overcomes limitations of these taggers. In addition, it was necessary to devote the Section 6 to highlight other annotation forms primarily the parsing and semantic analysis. Finally, we reach the conclusion in Section 7.

In Chapter 5, each Section is devoted to presenting with detailed information about the building procedures and characteristics of corpora developed during this thesis. Furthermore, they are compared to similar state-of-the-art corpora, stressing their significant contribution to the literature. Expressly, Section 2 is devoted to present Al-Mus'haf corpus, a corpus that includes the Quranic texts annotated with morphosyntactic information. Section 3 presents the OSIAN corpus, an Open Source International Arabic News corpus. It is a result of a collaborative project involving partners from Leipzig University in Germany. The purpose of this project is building an enormous collection of country-specific Arabic corpora. Furthermore, the MulTed corpus is introduced in Section 3. MulTed is a multilingual parallel corpus which is PoS tagged and sentence-aligned bilingually, with English as a pivot language. This corpus is compiled based on subtitles extracted from 1,100 TED talks. Finally, the Chapter is concluded in Section 5.

By way of conclusion, in Chapter 6, we summarize the key contributions of this thesis, the written and published papers issued from it as well as its limitations, and future research directions. It is worth mentioning that all Arabic scripts cited in this thesis are transliterated using Buckwalter transliteration[3].

[3] http://www.qamus.org/transliteration.htm

CHAPTER 2: Background and Literature Review

1. Introduction

According to EAGLES (the Expert Advisory Group on Language Engineering Standards) (Sinclair 2004), a corpus is a collection of naturally occurring samples of a language. These samples are selected and stored in electronic format; they are ordered according to external criteria to represent, as far as possible, a language as a source of linguistic research. Thus, the compilation process is to combine texts of both written and spoken language into a corpus. The design, compilation, and analysing corpora have led to the creation of a new scholarly field known as corpus linguistics. According to *McEnery* and *Hardie* (2011) corpus linguistics is a heterogeneous field, consensually an agreed set of methods and procedures for the exploration of language.

To make the following sections equally rich in both theoretical and practical aspects, a survey, that covers 100 of well-known and influential corpora, has addressed distinct stages in corpus building. Since the English language was the forerunner in corpus linguistics, it is obvious that 25% of the covered corpora in this survey are devoted to English. However, many other languages are catching up, implicitly considering English corpora as a global standard. Note that, information regarding the website addresses or DOI (Digital Object Identifier) for all data mentioned in this survey are given in Appendix "B". In Section 2, we present the early history of corpora including the first corpus to be developed and the first corpus to be given that name. The design criteria and corpus typology are described in Section 3. Section 4 provides information about well-known data sources, relevant compilation methods, and the most suitable formats for publishing corpora. Section 5 addressed the Arabic language features, followed by an overview of the major progress achieved in building Arabic corpora. Finally, we conclude this chapter in Section 6.

2. Origin and Earlier Corpora

Generally, the earlier corpora, those occurred before the 1960s, are called pre-electronic because they were not computerized and consisted of work in few areas. It is important to say that building and using corpora is not a totally new method or a scholarly field especially for linguists, but using the word "corpus" and establishing corpus building standards are the new events of the last decades. In fact, the roots of developing a corpus can be traced back to an Arabic lexicographer. We cite bellow some important moments in the history of building corpora:

- In the 8th century, *Al-Khalil bnu Ahmed Al-Farahidi*, an Arabic phonologist and lexicographer, assembled large collections of texts to develop the first Arabic dictionary and one of the earliest known dictionaries of any language, called "Kitab al-Ayn" (Versteegh 2014).
- The first concordance, completed in 1230, was produced based on the Bible (James 2015), it has been said that 500 monks engaged in its preparation.
- In 1897, the German linguist *Kaeding*, was able to publish the first known word frequency list based on word counting (Khorsheed et al. 2009). In doing so, he compiled a large corpus of German that contains 11 million words with the help of over five

thousand assistants over a period of several years needed to process the corpus (Bongers 1947; Kennedy 2014).

- In 1964, the first corpus to be given that name was developed, released, and known as the Brown Corpus (Francis and Kucera 1982; Hunston 2013). It consisted of texts amounting to over one million tokens after compiling a single year of publication (1961).
- During the 1970s, building corpora was characterized by its international aspect. Universities and research centres in Europe lunched several joint projects. The collaboration between the universities of **L**ancaster, **O**slo, and **B**ergen, which produced the **LOB** corpus of written British English in 1976, is perhaps the most prominent example.
- The real appearance of corpus linguistics was in the 1980s thanks to the widespread use of computers and access to machine-readable texts (O'Keeffe and McCarthy 2010; Kennedy 2014). Computers made it possible to collect much larger quantities of text and to process them much more quickly. As an example, at that time, the size of the Bank of English corpus was about 20 million tokens (Sinclair and others 1987).
- During the 1990s, the concept of a National corpus emerges. The first corpus of this kind was the British National Corpus (BNC) (G. Leech 1992b), which was begun in 1991 and released in 1994. With a size of 100 million tokens, the BNC was many times larger than any previous corpus.
- Finally, the web revolution paves the way for building corpora based on web content. Consequently, the size starts to escalate quickly to billions of tokens. For example, the GDELT Project[4] monitored 9.5 billion words of worldwide Arabic news over 14 months (February 2015 to June 2016) to make over a dataset of 6 million trigrams for the Arabic language.

3. Corpus Design and Corpus Typology

As seen in the previous section, building corpora was for a purpose and the texts were not haphazard collections but were made systematically. Consequently, most of them are used for different purposes. Unfortunately, for Arabic as for many other languages, most researchers have often built corpora that may only suit their personal objectives and for a specific time without considering the design criteria to create a well-defined corpus that meets the standards. As a result, the use of these corpora may lead to different results than expected because constructed corpora that are based on undefined design criteria, unidentified objective, or on technically unsuitable forms will be less used. Moreover, they will be neglected if the size is not enough to perform linguistic analysis or to train and evaluate NLP applications.

McEnery and Wilson (2001) claim that a corpus has four main characteristics:

1. sampling and representativeness;
2. finite size;
3. machine-readable form;
4. status as standard reference.

[4] https://www.gdeltproject.org/

In fact, a corpus is designed and compiled according to some explicit criteria defined by its developers in order to be representative samples of languages (G. N. Leech 1991). Further, the aim of these criteria is to identify at least what type of corpus is being constructed. In the following subsections, we introduce the fundamental guidelines of corpus building and corpus typology.

3.1. Design Criteria

Since the 60's, the design of corpora has been an obstacle for corpus linguists. However, in the 90's, basic guidelines have been set by relevant authors namely (G. Leech 1992b; T. McEnery and Wilson 1996; Sinclair 1996; Biber et al. 1998). Later, these guidelines were expanded by *Sinclair* (2005) in 10 fundamental criteria. In addition, *Leech* (1992a) made clear that corpora are not haphazard collections of textual material. Thus, a great care must be taken during the compilation process, otherwise the corpora developed will lead to results different from the expected. The corpus design usually starts from identifying the appropriate criteria, which means that the corpus likely seeks to be representative with respect to the phenomena under investigation (Ball 1994). If there are no specific criteria, the corpus should be designed for a general use to suit most corpus-based studies.

As mentioned, *Sinclair* (2005) formulates the overall instructions proposed by the previous authors in ten fundamental criteria to follow in the design and the compilation of a general corpus:

1. *The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.*
2. *Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.*
3. *Only those components of corpora which have been designed to be independently contrastive should be contrasted.*
4. *Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.*
5. *Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.*
6. *Samples of language for a corpus should, wherever possible, consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.*
7. *The design and composition of a corpus should be fully documented with information about the contents and arguments in justification of the taken decisions.*
8. *The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.*

9. *Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.*
10. *A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.*

Although the mentioned ten guidelines are core principles to design and compile a general corpus, they may not always suit every potential corpus builder. *Sinclair* (2008) himself notes that some of these guidelines can be difficult to uphold because of the nature of the language itself. However, the representativeness, balance, and homogeneity in the design process are necessarily idealistic. Alternatively, specialized corpora are designed relative to individual research aims, such as creating a dictionary, studying and providing analysis of the used language in a specific subject domain. However, it is advised to design a specialized corpus regarding Sinclair's guidelines, as ultimately a reliable or generalizable result can be derived from the analysis.

3.2. Corpus Typology

Success in corpus design leads to identifying the type of the corpus. Basically, identifying the types of corpora leads back to their primary uses in order to fulfil research needs. However, overlaps are inevitable in this classification. Next, we outline the main types of corpora based on one of the most relevant reference book in corpus linguistics (Lüdeling and Kytö 2008).

Historically, the first type, which has an international dimension, was the Brown family. As mentioned before, in the early 1960s, the Brown corpus was the first modern corpus for written American English. Then, during the 1970s the Brown corpus was part of speech tagged; the tagging was done automatically and was subsequently edited. Later, other corpora (e.g., LOB, Kolhapur (Shastri 1988), and the Australian Corpus of English (Collins and Peters 1988)) used the same sampling technique, the same text categories, and the same size as the Brown corpus. This kind of corpora is characterized by its representativeness, balance, and sampling. The texts were divided into an average of 15 categories, from which 500 samples of 2,000 tokens were then drawn proportionally from each category, totalling one million tokens.

The 1990s was the beginning of the concept of "National/Reference corpora". The purpose was to build general reference corpora to represent all the relevant varieties and the characteristic vocabulary of a national language of a country. For this purpose, several projects were launched to build large (at least 100 million tokens) and balanced corpora with regard to three criteria: domain (i.e., subject field), time (i.e., period of production), and medium (type of source). In addition, these corpora intended to cover, as much as possible, both written and spoken language. The BNC, as known, was the model followed by other languages such as German (Ćavar et al. 2000), Hungarian (Váradi 2002), Korean (Kim 2006), and Turkish (Aksan et al. 2012).

Most reference corpora are essentially static samples and achieve a steady size. Unlike them, a new type of corpora emerges that was more dynamic, and constantly updated to track rapid language change by tracking new words, new uses or meanings of old words, and words falling out of use. This type is called "monitor corpora". The best example of this type is the Bank of English (BoE) corpus (Järvinen 1994). The BoE corpus was designed to represent the international English language in use in present-day. It consists of 75% of written texts come from several sources such as newspapers, fiction books, and websites; while the remaining 25%

are spoken data like transcripts of television and radio broadcasts. The BoE was initiated in 1991 and is being updated every two years since 2000; up to date, it contains 650 million words. Similarly, another corpus has been designed in 1998 to be a reference corpus of contemporary Italian. Then, it was updated every two years in order to build a monitor corpus which presently includes 130 million words. The corpus referred to here is the CORIS corpus (Favretti et al. 2002).

There are other types that are worth mentioning: the synchronic and the diachronic corpora. Synchronic corpus is a reliable basis for comparing language varieties. It consists of written and spoken data produced during a specific period in countries in which the same language is a first or official language. A typical example of this type is the International Corpus of English (ICE) (Nelson et al. 2002), which consists of 23 sub-corpora of one million words each, covering data produced in English during 1990-1994 in some countries like Great Britain, the USA, Australia, India, Nigeria, and Hong Kong. On the other hand, to explore a language change from a historical perspective, a diachronic corpus is the ideal choice. It contains texts from the same language gathered from different time periods. As examples, the historical part of the Helsinki corpus (Rissanen and others 1993), which consists over 1.5 million words of English, dating from the 8th to the 18th centuries, and the KACST Arabic corpus (Al-Thubaity 2015) which comprises over 731 million words from pre-Islam until 2013 (more than 1,500 years). The corpus aims to represent the Classical and Modern Standard Arabic language and the transition between them.

4. Corpus Compilation

Building a corpus is not a trivial task, which might be reflected by the fact that most of the early research focused on a small number of well-formed corpora in very few languages. Here, we present most well-known data sources, compilation methods, and suitable formats for building corpora.

4.1. Data Sources

Based on the previous design criteria, corpora builders should identify the corpus source genres and size. The selection and finding of suitable resources are much more complicated. For example, as stressed by *Lüdeling* and *Kytö* (2008), the components of general corpora typically are representative of various genres, whilst specialized corpora can be limited to highlight only one genre or a family of genres. Regarding size and balance, *Lüdeling* and *Kytö* claim that it is not always possible to collect data in similar (or even sufficient) quantities for each text category represented in the corpus; this is often the case with historical corpora. Moreover, one of the biggest challenges all corpora builders encounter is the lack of public resources and copyright.

There are several sources that can be used to build corpora. Corpora, on one hand, may consist of a single book like the ones developed and used by *Baneyx et al.* (2007) to build an ontology of pulmonary diseases and by *Liua et al.* (2016) to build common-sense knowledge enhanced embeddings to solve pronoun disambiguation problems. On the other hand, corpora can be developed using several books like the Shamela corpus (Belinkov et al. 2016), or editions of a particular newspaper (Maamouri et al. 2013). Recently, corpora builders, in particular individuals, use the web to build very large corpora in a short time and with low cost (Nakov 2014). However, we must be cautious while building a balanced corpus in which data of a language must be drawn from a wide range of sources.

Regarding our survey (Zeroual and Lakhouaja 2018), we proposed 9 well-known sources (Books, Web, Magazines, Newspapers, Records, Video subtitles, Official prints, Human-generated, and Dictionaries), offering the possibility to have multiple redundant or adding new sources that are not listed. The results are presented in Table 2.1.

Table 2.1 The used sources to build corpora

Sources of corpora	**Nb. of corpora**
Books	11
Books, Official Prints, Newspapers, Magazines	16
Dictionaries	1
Human-generated (e.g., summaries from original documents)	4
Newspapers	12
Records	2
Video subtitles	3
Web	22
Web, Books, Official Prints, Old Manuscripts, Newspapers, Magazines	22
Web, Human-generated	3
Web, Newspapers, Magazines	4
Total	**100**

As seen in the Table, sometimes corpora builders use more than one source to compile their corpora, therefore, we have 11 cases. For instance, the first row of the table says that 11 corpora were build using only one source such as the Negra corpus (Brants et al. 2003) which consists of German newspaper texts taken from the *Frankfurter Rundschau* and the arTenTen corpus (Arts et al. 2014) which is a web-crawled corpus of Arabic. On the other hand, 22 corpora, including the BNC corpus, have been built using different sources such as Web content, books, official prints, newspapers, and magazines. Furthermore, the most used sources are the Web content and books as 51 of corpora are compiled based, totally or partially, on the web content and 49 corpora on books.

4.2. Compilation Methods

Understandably enough, corpus compilation is time-consuming and often difficult to undertake depending on the corpus size and the methodology adopted. Through the survey we conducted, it was possible to collect information about different compilation methods used to build 100 corpora in relation to their size and the average time consumed during the compilation process. The purpose of this study is not to determine the better compilation method. Instead, we would like to know what relevant methods are used and how long they took to complete the procedure. Table 2.2 exhibits a summary of the compilation methods mentioned in the survey. We can observe that corpora builders usually rely on three major methods:

1. The first and the oldest method ever, the manual method.

2. The inevitable result after the appearance of computers, the automatic method.
3. The latter is not completely accurate because the crawled data are often duplicated on the Web and need to be cleaned, filtered, converted into the right format, and annotated. Then, a manual edition is subsequently performed to proofread the previous processes. This is the semi-automatic method.

Table 2.2 Compilation methods of corpora

Compilation methods	Corpora size (Nb. of tokens)	Nb. of corpora	Average time consumed
Manual (28%)	<= 100K	5	2 years
	<= 1M	8	3 years 6 months
	<= 10M	7	4 years 6 months
	<= 50M	4	7 years
	<= 100M	1	3 years
	<= 500M	2	3 years
	<= 1Bn	1	7 years
Semi-automated (40%)	<= 100K	1	2 years
	<= 1M	8	3 years
	<= 10M	6	4 years 6 months
	<= 50M	3	5 years
	<= 100M	3	2 years
	<= 500M	10	5 years
	<= 1Bn	3	6 years
	> 1Bn	6	10 years 3 months
Automated (25%)	<= 10M	1	3 years
	<= 50M	1	4 years
	<= 100M	6	5 years 8 months
	<= 500M	5	3 years 10 months
	<= 1Bn	3	2 years 6 months
	> 1Bn	9	3 years
Crowdsourcing (6%)	<= 100K	2	1 years 9 months
	<= 1M	2	7 months
	<= 50M	1	8 years
	<= 1Bn	1	Since 2001
Gamified approach (1%)	<= 100K	1	43 days

Concerning the crowdsourcing method, it is catching up. Crowdsourcing was the result of the advancement of digital technology and the Internet in the mid-2000s. It is mainly an online

sourcing model in which individuals or organizations use contributions from online communities for problem solving and resources production (Brabham 2013). One of the main platforms used in crowdsourcing is Amazon's Mechanical Turk[5] (e.g., (El-Haj et al. 2015; Habernal and Gurevych 2016)). The last observed method is the gamified approach or "Gamification". It is a new-coming method to NLP (Zeroual et al. 2017), it was first mentioned in 2003 and start to be used in literature in 2010 (Lund and O'Regan 2016) (e.g., (Tiam-Lee and See 2014)). By definition, gamification is the use of game design elements in non-game contexts to increase users' motivations towards given activities (Deterding et al. 2011). Though gamified approaches are based on the same strategy as crowdsourcing, in the latter, the participants receive money to increase motivation; whereas gamification incentivizes them by providing an experience of entertainment.

It can be supposed that producing a corpus with a considerable size and variety requires several years of efforts regardless of the used method. In the context of NLP, the first and the most way to collect meaningful and high-quality of data is to hire expert linguists to manually build or annotate corpora; however, it takes time, and costs money. The survey shows that 86% of manually built corpora contain less than 50 million words and it takes more than four years to complete them; whereas, 88% of automatically built corpora with a size that varies from 100 million to billions are completed in less than the same period.

Furthermore, the automatic methods marked a considerable progress lately, mainly due to the recent advances in deep learning technologies (e.g., deep neural networks), especially when it comes to dealing with very large data and access to information. The "Intelligent Personal Assistants" like Apple Siri, Google Assistant, Microsoft Cortana, Amazon Echo, etc. are certainly sufficient examples of the significant success achieved using these methods.

The semi-automated methods are a combination of both manual and automated methods. The aim is to raise the quality of data within a feasible/reasonable time and cost. In doing so, at first stage, the data are collected or annotated automatically and later edited by experts. It is worth mentioning that the shortest time consumed for building a corpus is achieved by the gamified approach. This method can be promising, especially it combines some key features of the other methods. For example, it is fast, as well as automated methods, and provides satisfactory quality results. Further, it is based on the same strategy as crowdsourcing, but its cost does not scale with data size.

4.3. Corpus Format

Basically, corpora differ from other electronic representations such as archives and databases. Archives are normally unstructured repositories of texts, whereas databases are collections of an entire population of data. The latter are designed to facilitate data entry and retrieval. Though corpora are a subset of databases, they are designed and compiled according to some explicit criteria as illustrated before.

Text encoding or markup is one of the main tasks after corpus compilation. Typically, corpora consist of electronic versions of texts taken from various sources. Therefore, a confusion may arise due to different codes used for markup. Since the 1980s, the Standard Generalized Markup Language (SGML) has become increasingly accepted as a standard way of encoding

[5] https://www.mturk.com

electronic texts. Using SGML is considered as a basis for corpus preparation; it facilitates the portability of corpora, enabling them to be reused in different contexts on different equipment, thus saving the cost of repeated typesetting. Since SGML can be complex for some corpus builders and users, an Extensible Markup Language (XML) was derived from SGML that contains a limited feature set to make it simpler to use.

5. Review of Arabic Corpora

This section addresses the Arabic language features and its linguistics specification primarily those considered as challenges in NLP. Further, a detailed overview of the major progress achieved in building Arabic corpora is given.

5.1. The Arabic Language

Arabic is an interesting language and a fruitful area of research for corpus linguistics, as much as it is a challenging language for existing NLP applications because of its characteristics. The Arabic lexicon and grammar are deeply rooted and well established a long time ago. The morphology of Arabic differs in the structure of affixes from Indo-European languages (Gharaibeh and Gharaibeh 2012). Arabic is a rich Semitic language and highly productive both derivationally and inflectionally (Alsaedi et al. 2016). These two complex paradigms are based on the interaction between roots and patterns which have intrigued lexicographers and morphologists for centuries. Moreover, a word can represent a whole sentence through sequential concatenation. For example, the Arabic word "أَنُلْزِمُكُمُوهَا" from the 28th verse of chapter 11 (sūrat Hud) <AanulozimukumuwhaA> means in English "Should we compel you to accept it".

Basically, the Arabic language consists of three main categories (Al-Dahdah 1989): Noun "اسم" <Asm>, Verb "فعل" <fEl> and Particle "حرف" <Hrf>. In addition, each one of these categories has dozens of subcategories (Zeroual, Lakhouaja, et al. 2017). According to Habash (2010), the diacritics (short vowels) in Arabic script are optional. As a result, Arabic words can be written either fully diacritised, partially diacritised, or non-diacritised. The omission of diacritics in written modern standard Arabic has posed some difficulties to several automatic processing systems (Chennoufi and Mazroui 2016); when these vowels are omitted, they are left for the reader to infer, knowing that the vowels can encode grammatical category or feature information. In addition, most Arabic roots consist of three consonants and the vowels add grammatical information when attached to these consonants. What's more, it is estimated that the average number of possible part of speech tags for a word in most languages is 2.3, whereas in modern standard Arabic is 19.2 (Farghaly and Shaalan 2009). For example, the three consonants "كتب" ktb can stand for the verb "كَتَبَ" <kataba> "he wrote", or for the plural noun "كُتُبٌ" <kutub> "books", among other part of speech tags.

The free word order nature in Arabic sentences is another feature that makes parsing one of the most difficult tasks. I.e., we could easily change the order between the subject and the verb without the need for an agreement between one another in number (singular or plural). For instance, the following sentences are both correct in Arabic: "الأَوْلاَدُ يَلْعَبُونَ" and "يَلْعَبُ الأَوْلاَدُ" which literally mean "The boys play" and "plays the boys", respectively. These Arabic language features are generally the most common challenges faced by researchers in the Arabic corpus linguistics and NLP fields.

5.2. Overview on Arabic Corpora

In this section, we provide an overview of the state-of-the-art of Arabic corpora. There has been over the past few years a tremendous growth in interest and activity in Arabic corpus building and analysis area. Besides, encouraging works have been undertaken recently. Next, we list a number of relevant and recent corpora focusing on their objectives and scope. The listed corpora are mainly classified based on their target language and mode, and we mention some of their characteristics such as their designated purpose, availability, size, text domain, and the presence of annotations.

5.2.1. Quranic Corpora

The Arabic language was codified primarily in the Quran (Watson 2002) and based on the language of the western Hijazi tribe of Quraysh, with some interference from pre-Islamic poetic *koiné* and eastern dialects. The Quranic scripture is used to guide the lives of 1.6 billion Muslims worldwide and they use it to perform their daily prayers (Yassein and Wahsheh 2016). The Quran is the finest piece of literature in the Arabic Language, and the number of non-Arabic speakers that learn the Arabic language with the objective of understanding the Quran is rapidly increasing. The Quran contains over 77 thousand words, it is divided into 114 chapters where each chapter is divided into verses, adding up to a total of 6,243 verses. Some relevant corpora are created from the original text of the holy Quran, namely:

- Quran Corpus of Haifa (Dror et al. 2004): This corpus has been built using an automatic morphological analysis on the Quranic text. However, the work is not complete, it remains manually unverified and has multiple possible analyses for each word in the final published data set. Considering a random sample, the authors of the Haifa corpus estimate the final accuracy of annotation using the F-measure at 86%. Further, approximately 40% of the roots in the Haifa corpus are missing and the words' lemmas are not given.
- The Quranic Arabic Corpus[6] (Dukes and Habash 2010): It is an online-annotated corpus with multiple layers of annotation including morphological segmentation, part of speech tagging, syntactic analysis using dependency grammar and a semantic ontology. Despite this corpus being manually verified, it has some problems on the level of lemmas and roots, and has insufficient grammatical information. Furthermore, the patterns are not given.
- QurAna corpus (Sharaf and Atwell 2012a): In this corpus, only the personal pronouns are tagged with antecedent information (over 24,500 pronouns). These antecedents are maintained as an ontological list of concepts. The Quranic Arabic Corpus was used to identify the targeted segments that contain pronouns, and for each pronoun, the starting and ending IDs of the text span that represents antecedents were recorded manually through forms developed using the PHP scripting language.
- QurSim corpus (Sharaf and Atwell 2012b): It is an annotated corpus where semantically similar or related verses are linked together. With the help of domain experts, the authors adopt the same methodology of *Ibn Kathir*, a Muslim scholar who is known for his classic book of Quran commentary (or Tafsir in Arabic). In fact, the principle of this method is to link two verses if one of them was cited while commenting on the other. The size of the dataset is over 7,600 pairs of related verses and the authors claimed that this dataset could be

[6] http://corpus.quran.com/

extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs.

- The Boundary-Annotated Quran Corpus (Sawalha et al. 2014): Unlike the other Quranic corpora, the words in this corpus are tagged with prosodic and boundary annotation rather than morphological or syntactical annotation. It was built by gathering and tracking boundary stops from the "Tanzil Quran project"[7], the part of speech tags from the Quranic Arabic Corpus, and the prosodic annotation scheme from Tajwid (recitation) mark-up in the Quran.
- Qurany[8]: the Quranic text is augmented with an ontology or index of key concepts that were imported from "Mushaf Al Tajweed", a recognized expert source which is compiled by Dr. Mohamed Habash, Director of the Islamic Studies Centre in Damascus, published by Dar Al-Maarifah in Syria and authenticated by the Al-Azhar Islamic Research Academy in Egypt. The Qurany allows users to search in the Holy Quran for abstract concepts via an ontology browser. Users can use this browser to identify a precise concept among nearly 1200 concepts and find the related verses to this concept; yet, the corpus includes 8 variant English translations.
- Al-Mus'haf corpus (Zeroual and Lakhouaja 2016): It is an enriched corpus with morphosyntactical information. The process of building this corpus consists of a semi-automatic technique by using "AL-Khalil Morpho Sys2" (Boudchiche et al. 2016), then manual processes. The corpus has 1770 roots, vowelled patterns for each stem and lemma, over 100 part of speech tags used, and true lemma (1554 patterns).

5.2.2. Classical Arabic Corpora

The Classical Arabic (CA) is the form of the Arabic language particularly used in literary texts and applied on the academic and religious levels. The Quran is considered to be the highest form of CA texts. The amount of published CA texts is higher than the texts published in Modern Standard Arabic. Consequently, the free and large linguistic resources published by the Arabic corpus linguistic community are available in CA. For instance:

- The King Saud University Corpus of Classical Arabic (KSUCCA) (Alrabiah et al. 2013): the corpus contains 50M+ words. The data of the corpus includes only pure CA texts, the resources dated back to the period of the pre-Islamic era until the end of the 4th Hijri century (equivalent to the period from the 7th to early 11th century CE). The corpus covers six broad genres which are most of the topics that were popular in that period. These are: Religion, Linguistics, Literature, Science, Sociology, and Biography. The major resources of the corpus were extracted from the Shamela[9] library. Recently, the corpus has been tagged using MADA+Tokan (N. Habash et al. 2009) with a tagset that consist of 41 basic tags where the estimated accuracies are: 87.80% for lemmas, 84.90% for roots, 83.40% for part of speech tagging, 89.90% for masculinity and femininity, and finally, 90.10% for singularity and plurality. KSUCCA[10] is available for download.

[7] http://tanzil.net/

[8] http://quranytopics.appspot.com/

[9] http://shamela.ws/

[10] http://ksucorpus.ksu.edu.sa

- Shamela (Belinkov et al. 2016): It is a large-scale historical Arabic corpus from diverse periods of time (from the 7th century to the modern era). The corpus is drawn from the Shamela library, it is a voluntary project accomplished by the cooperation between the site's owners and the Alrawdhah[11] Cooperative Office for Call and Guidance. The corpus is cleaned and organized with a metadata information in a semi-automatic process. Moreover, the entire corpus is processed with Madamira (Pasha et al. 2014a), a state-of-the-art morphosyntactical analyser and disambiguator. The result is a full analysis per word, including tokenization, lemmatization, part-of-speech-tagging, and various morphological features. The corpus contains over 6,000 texts, totalling around 1 billion words, of which 800 million words are from dated texts and the remaining texts are automatically dated by building a 5-gram language model with Kneser-Ney smoothing, using the SRILM toolkit (Stolcke et al. 2011).
- Tashkeela (Zerrouki and Balla 2017): It is a recent corpus of Arabic raw and diacritized texts that contains over 75 million of fully vocalized words obtained from 97 Islamic books filtered from 7079 books from Shamela Library. These classical books present 98.85% of corpus data, while 1.25% are data collected from 20 modern books and texts crawled from Internet websites such as Aljazeera Learning Arabic[12]. Tashkeela corpus is available for download from its project website[13].

5.2.3. Modern Standard Arabic Corpora

The Modern Standard Arabic (MSA) is the form used in contemporary scholarly published works as well as in the media. MSA does not differ from CA in morphology or syntax, but richness of stylistic and lexis usage is apparent in Classical works. Most researchers on Arabic corpus linguistics have concentrated on MSA, and they have an aspect of making their works publicly available which can be used for different purposes. In 2010, the Mediterranean Arabic Language and Speech Technology (MEDAR) conducted a survey[14] to list the projects carried out for developing Arabic language resources. Unfortunately, the list produced by this survey is no longer updated. Another and probably the very recent survey (Zaghouani 2017) is conducted to identify the list of the freely available Arabic corpora and language resources. The survey published an initial list of 66 sources. A few major examples are mentioned here to give an idea of the variety of what is available.

- arabiCorpus[15]: It is a free online set of Arabic corpora developed by Parkinson (2012). The corpus contains about 146.000.000 tokens from written and spoken materials. Besides, arabiCorpus includes different MSA as well as premodern resources such as Newspapers, modern literature, Egyptian colloquial, and religious books. however, it belongs to the MSA category since the newspapers accounting for over 90% of the total size of the entire corpus.
- Open Source Arabic Corpora (OSAC) (Saad and Ashour 2010): It is a collection of largest free accessible raw corpora and a large number of web documents extracted from over 25 Arabic websites using the open source offline explorer, *HTTrack*. The compilation of corpus

[11] http://www.arrawdah.com
[12] http://learning.aljazeera.net
[13] https://sourceforge.net/projects/tashkeela/
[14] http://www.medar.info/MEDAR_Survey_III.pdf
[15] http://arabicorpus.byu.edu/

included converting html/xml files into UTF-8 encoding using "Text Encoding Converter" by *WebKeySoft* and removing the html/xml tags. The corpus contains about 113 million words and covers several topics: economy, history, education, religion, sport, health, astronomy, law, stories, and cooking recipes.

- Alwatan-2004 (Abbas et al. 2011): The main purpose of compiling this corpus is to evaluate topic identification methods, but it is suitable for other Arabic NLP tasks. The corpus contains 20,291 documents, corresponding to 10 million words, and is organized in six topics (religion, economy, local news, international news, and sport). These documents were downloaded from the Saudi newspaper Alwatan[16]. The texts were prepared by removing punctuation marks, digits, and Stop-List words. For the next stage of treatment, the corpus was tagged via the KALIMAT Corpus (El-Haj and Koulali 2013) using the Stanford Arabic part of speech tagger (Toutanova et al. 2003) (33 basic tags).
- El-Haj (2015) created several resources (e.g., KALIMAT, EASC, and ABMC) for those working on computational methods to analyse and study languages primary Arabic. These resources are articles collected from the Arabic language version of Wikipedia and two Arabic newspapers (Alrai[17] and Alwatan[9]), and human-generated extractive summaries of those articles. A group of students was asked to search for Wikipedia and select articles within ten given subject areas (art and music, the environment, politics, sports, health, finance and insurance, science and technology, tourism, religion, and education). El-Haj et al. used Amazon's Mechanical Turk to recruit appropriately skilled human participants to create the gold-standard summaries of the collected articles. The final corpus includes a total of 153 documents, containing 18,264 words and each document consists an average of 380 words, with a minimum of 116 words and a maximum of 971 words.

Several MSA corpora were designed and built for more specific purposes but they can be used for other purposes if the relevant research question(s) can be answered. The Corpus of Contemporary Arabic (CCA[18]) (Al-Sulaiti and Atwell 2006) and the Arabic Learner Corpus v2 (ALC) (Alfaifi et al. 2014) are generally compiled for language teaching and learning research. The ALC is a written and spoken MSA corpus developed at Leeds University between 2012 and 2013. The data were produced by 942 learners of Arabic, from 67 different nationalities studying at pre-university and university levels in Saudi Arabia. The available corpus comprises 282,732 words stored in TXT and XML formats, hand-written sheets which are in PDF format as well as the audio recordings which are available in MP3 format. The CCA corpus is designed to resemble the American National Corpus, it is also a written and spoken MSA corpus collected from 1990s up to 2005 and covers several topics; For the written part, the six major topics were fiction, arts, science, business, miscellaneous; and for the spoken part the data were derived from TV, Radio, Conversation. Concerning the resources selection phase, the authors carried out a survey of language teachers and language engineers to get their opinions on the texts that might be of use to them. As a result, they managed to compile a corpus of 1 million words. Finally, It may deserve to mention here, the Qatar Arabic Language Bank (QALB), an ongoing project to build a large error-annotated corpus of Arabic text with manual corrections (Zaghouani et al. 2014; Rozovskaya et al. 2015). The QALB corpus will be beneficial for corpus-based studies of

[16] http://www.alwatan.com.sa

[17] http://alrai.com/

[18] http://www.comp.leeds.ac.uk/eric/latifa/research.htm

errors and for design and develop Arabic automatic spelling and text correction tools. The corpus includes texts gathered from online user comments written to Aljazeera articles, it also includes texts produced by learners of Arabic as a foreign language. Next, these texts were processed with the morphosyntactical analyser Madamira. The texts are then manually annotated for errors. The errors include spelling, punctuation, word choice, morphology, syntax, and dialectal usage.

5.2.4. General Corpora

The general corpora can be used for general corpus-based studies or for inferring the general patterns of the Arabic language for NLP applications. Typically, they include a large number of data targeting more than one language and mode (e.g., MSA and CA). However, this category of corpora is rare and usually not available for free. As examples, we state:

- arTenTen (Arts et al. 2014): It is a member of the TenTen Corpus Family (Jakubíček et al. 2013). The arTenTen is a web-crawled corpus of Arabic, it was crawled using Spiderling (Suchomel et al. 2012) gathered in 2012. The arTenTen corpus is partially tagged: one sample of the corpus comprises roughly 30 million words that were tagged using the Stanford Arabic PoS tagger; another sample containing over 115 million words that were tokenised, lemmatised, and part of speech tagged using MADA system. The arTenTen comprises 5.8 billion words but it can only be explored by paying a fee via the Sketch Engine website[19].
- King Abdulaziz City for Science and Technology (KACST) Arabic corpus (Al-Thubaity 2015): It comprises over 731 million words from pre-Islam until 2013 (more than 1,500 years). The corpus aims to represent the two main forms of Arabic language (CA and MSA) and the transition between them. The resources were collected mainly from all Arab countries, but also Arabic publications from other regions. The KACST includes a wide diversity of texts content covering 10 mediums namely Old Manuscripts, Books, Newspapers, Magazines, Curricula, University Theses, Websites, Refereed Periodicals, Official Prints, and News Agencies. Each text has been further classified more specifically into 80 domains (e.g., Islamic and Arabic Poems) and 481 topics (e.g., Hadeeth, love, and wisdom). In order to allow the corpus users to study the Arabic language and its many varieties in both general and specific ways and across many different levels, thereby allowing for more accurate language models to be constructed, the following metadata were assigned to each text: title, year of publication, time period, author name and gender, region, medium, domain, and topic.

5.2.5. Dialectal Corpora

The colloquial Arabic dialects are the form of Arabic used in everyday oral communication. They differ significantly in each Arab country; nevertheless, most of them are mutually intelligible. Dialectal Arabic varieties, notably Egyptian (Maamouri et al. 2014) and Gulf Arabic (Khalifa et al. 2016), have lately received some attention and have a growing collection of resources that include annotated corpora (Cotterell and Callison-Burch 2014), a neural architecture for Dialectal Arabic Segmentation (Samih et al. 2017), sentiment analysis (Mdhaffar et al. 2017), and morphological analysers (Salloum and Habash 2014; Khalifa et al. 2017). Additionally, the first project to build a multidialectal Arabic parallel corpus (Bouamor et

[19] https://www.sketchengine.co.uk/

al. 2014) was launched four years ago. The corpus contains a collection of 2,000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian, Syrian Arabic, and English. Similarly, the Parallel Arabic Dialect Corpus (PADC) was created from 6,400 transcribed sentences that came originally from two Algerian dialects, then, the sentences were translated into Tunisian, Palestinian, and Syrian dialects considering MSA as a pivot language (Harrat et al. 2017).

5.3. Multilingual Corpora Including Arabic Language

The Arabic language has been included in relevant pioneering multilingual corpora, such as the open source parallel corpus (OPUS[20]) (Tiedemann 2012), which is the largest collection of freely available parallel corpora in more than 90 languages and includes data from several domains. This corpus comprises over 40 billion tokens in 2.7 billion parallel units (aligned sentences and sentence fragments). The largest sources of OPUS are legislative and administrative texts (mostly from the European Union and associated institutions), translated movie subtitles, newspapers, and localization data from open-source software projects. Arabic is one of the top languages covered by the OPUS sub-corpora that comprise well over 100 million tokens. For instance, the MultiUN corpus (Multilingual Corpus from United Nation) (Chen and Eisele 2012), which is available in all six official languages of the United Nation plus the German language, comprises 271.5 million Arabic tokens. Table 2.3 provides more statistics about MultiUN corpus.

Table 2.3 Statistics about the MultiUN corpus

Languages	Files	Tokens	Sentences
Arabic	68,870	271.5M	11.1M
Dutch	4,034	6.7M	0.2M
English	100,373	443.5M	17.2M
Spanish	5,683	30.1M	1.0M
French	90,826	474.3M	14.9M
Russian	81,258	328.5M	13.9M
Chinese	69,360	83.1M	10.9M

Except the OPUS sub-corpora, the Arabic language is covered by a small number of bilingual and multilingual corpora such as the tiny Arabic-English parallel corpus (10K sentences) used to build an Arabic stemmer based on statistical machine translation using an English stemmer (Rogati et al. 2003). A similar Arabic-English parallel corpus has been adopted to handle the word translation disambiguation (Ahmed and Nürnberger 2008). In addition, a multilingual named entity corpus for Arabic, English, and French has been developed based on comparable newswires from the "Agence France Presse" covering the period 2004-2006 (Mostefa et al. 2009). Finally, a free Arabic-English parallel corpus has been built within the project MEDAR (Maegaard et al. 2009) which has been running from 2008 to 2010.

[20] http://opus.nlpl.eu/

The Arabic language is also present in multilingual corpora whose data are based on video's subtitles. For example, the AMARA corpus (Abdelali et al. 2014), a parallel corpus of educational video subtitles, multilingually aligned for 20 languages including Arabic. The data of this corpus are collected in cooperation with Amara platform[21]. 3000 videos have available subtitles in at least six languages and 1000 videos have available subtitles in 25 languages. A similar project[22] called "WIT^3" (Cettolo et al. 2012), an acronym for Web Inventory of Transcribed and Translated Talks. It is a collection of lecture translations that have been automatically crawled from the TED talks[23] in 109 languages. The purpose of this project is to support the machine translation evaluations campaigns of the International Workshop on Spoken Language Translation (IWSLT) (Paul et al. 2010). As of October 2011, 17 thousand transcripts corresponding to translations of around 1000 talks have been collected. The "WIT^3" comprises 2.4 million Arabic tokens.

6. Conclusion

Before building a corpus, some questions should be asked such as: Which data will be collected, speech, writing data or both? What are the time periods in which the data was produced? What is the suitable size for the corpus? How to balanced? What research questions is the corpus trying to answer?

So, perhaps the most important stage of all is the very first one, design. Because, without a solid design, everything else is likely to go wrong. In addition, it is important to bear in mind that the purpose for which a corpus is compiled influences its design. Consequently, success in corpus design leads to identifying the type of the developed corpus. Also, what distinguishes corpora from each other are the sources selected to collect data and the type of methods used during the compilation and annotation processes. Many relevant corpora were covered in this chapter that, to some extent, represent the literature of Arabic corpus linguistics and reflect the major progress achieved in this field.

The goal of this chapter has been to help readers better understand the data design, collection, annotation, and analysis procedures for corpus building by sampling from already published corpora of the most prominent projects of resource rich–poor languages. To make the survey rich in information, we targeted well-known corpora in the literature and those publicly available. Moreover, we covered many languages on purpose to ensure that the survey is as balanced as possible. In the following chapter, we move forward to describe the next stages in corpus building, data processing and annotation.

[21] http://amara.org

[22] https://wit3.fbk.eu/

[23] http://www.ted.com/

CHAPTER 3: Data Processing

1. Introduction

The main goal of text processing (also known as text handling) is to structure the corpus data into a suitable representation. Expressly, text processing is generating more representative terms linked to the original words. Morphological analysis is one of the most important techniques used in corpus processing. It deals with inflection and derivation paradigms especially for morphologically rich languages such as Arabic. Stemming and lemmatization are basic morphological analysis for NLP, and they affect directly the performances of subsequent analysis. Since both these tasks share a common goal of reducing a word to its base, they are sometimes used interchangeably. However, as used in Arabic NLP, the terms refer to different processes and differ slightly from its corresponding western languages. In this Chapter, all these concepts are clarified with numerous examples and experiments on the Arabic language. Also, relevant state-of-the-art methods and tools are presented, evaluated, and compared to our proposed methods.

2. Stemming

Basically, stemming is probably the first and main process used for handling morphologically rich languages, such as Arabic. The goal of stemming is to reduce inflected and derived words to their base (root or stem). It is an essential task in several fields primarily NLP and Information Retrieval (IR). Considering that Arabic is mainly dependent on roots and patterns to generate words, it is recommended that Arabic stemmers should be developed based on the interaction between roots and patterns to gain more efficiency. In this section, we present a hybrid system composed of two phases: a rule-based algorithm and a probabilistic model. In the first phase, the algorithm is based on root-pattern interaction; yet, lexicons are involved to help solving ambiguity issues. In the second phase, to identify the correct stem according to the context, we implement the Hidden Markov Models, smoothing techniques to circumvent the problem of missing transitions words, and the Viterbi algorithm to select the optimal solution. Then, we highlight the performance of the developed stemmer via various experiments on both MSA and CA.

2.1. Terminology and Classification of Stemmers

2.1.1. Terminology

Some terms must be clarified before going any further. In this section, we provide a clarification of basic concepts to avoid any confusion about the Arabic stemming. The following definitions are extracted from relevant works that address the Arabic morphological analysis (Attia et al. 2016; Singh and Gupta 2016; Zeroual et al. 2015):

- **Affixes:** They can be concatenated to a root to indicate grammatical features such as gender, verb tense, number, and person. The Arabic affixes are 10 letters: س, أ, ل, ت, م, و, ن, ي, ه, ا. We can collect them in a single word "سألتمونيها" <s>ltmwnihA>. There are three types of affixes depending on their position:

- Prefixes are those attached at the beginning of the word. For example, when referring to an event in the present, we have to add the prefix "ي" 'y' before the root. Also, since Arabic words cannot begin with a consonant cluster, the prefix "ا" 'A' is added at the beginning. An example is "اجْتَهَدَ" <Ajotahada> 'he worked hard'.
- Suffixes are attached at the end, for example, suffixes that display grammatical relations associated with female beings "مُدَرِّسَة=مُدَرِّس+ة" <mudar~isap= mudar~isa+p> 'Teacher'.
- Infixes are found in the middle of the word. For instance, "كاتب=ك+ا+تب" <kAtb = k+A+tb > 'Writer'.

- **Clitics:** They are linguistic units attached to a stem. They are pronounced and written as affixes but they are grammatically independent (Alotaiby et al. 2010). For example, prepositions, conjunctions, and pronouns. There are two types of clitics depending on their position:
  - Proclitics are attached at the beginning of the stem. For example, "فَنَامَ" <fanaAma> 'And he sleeps'.
  - Enclitics are attached at the end like the attached pronouns (e.g., "كتابه = كتاب+ه" <ktAbh = ktAb+h> 'His book').

  Figure 3.1 shows a construction of a model of clitics attached to a stem.

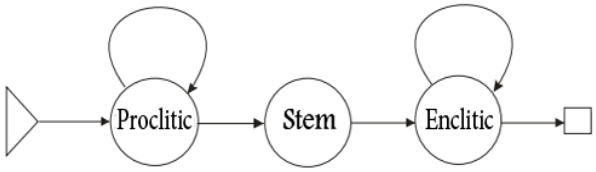

Figure 3.1 Model of clitics attached to a stem

Table 3.1 shows some examples of clitics (proclitics, enclitics) with a length ranging from 1 to 6 characters.

- **Stem:** It is the word without clitics (Larkey et al. 2007). A stem is a result of combining a root with inflectional affixes to indicate grammatical features. I.e., it conflates a set of surface words that share those features;
- **Root:** It is a sequence of mostly three consonantal radicals, which together signify some abstract meaning (Fabri et al. 2014). For example, the words "كَتَبَ" <kataba> 'to write', "كَاتِب" <kaAtib> 'writer/author', "مَكْتُوب" <makotuwb> 'written/letter', "مَكْتَب" <makotab> 'office' and "مَكْتَبَة" <makotabap> 'library' all share the same root morpheme "كتب" <ktb> 'writing-related'. About 11,347 roots are distributed as follows (Ababneh et al. 2012):

- 115: Two-character roots (e.g., "من" <mn>).
- 7198: Three-character roots (e.g., "كتب" <ktb>).
- 3739: Four-character roots (e.g., "دحرج" <dHrj>).
- 295: Five-character roots (e.g., "سفرجل" <sfrjl>).

Table 3.1 Examples of clitics

		Length	Example of Clitics	Example of surface words
Clitics	*Proclitics*	1C	**ب - س**	سأرى "I will see" – برفق "lightly"
		2C	**وب - ال**	وبفعل – "today" اليوم "and due to"
		3C	**بال - أفس**	بالفعل "in fact" – أفستذهب "are you going then?"
		4C	**أولل - وبال**	أوللرّجل – "does the man have?" وبالقيام "and by doing"
		5C	**أوكال - أفبل**	أوكاليوم – "is today as?" أفبالعنف "do with violence?"
	Enclitics	1C	**ك - ي**	كتابك "your book" – مدينتي "my city"
		2C	**كن - ها**	بيوتكن "your houses" – مدرستها "her school"
		3C	**وهن - كما**	رأيتموهن "you saw it" – أخذكما "he take you"
		4C	**نيهم - وناه**	أعطيتمونيهم "you gave them to me" – أعطيتموناه "you gave it to us"
		5C	**نيهما - كماهن**	أعطيتمونيهما "you gave both of them to me" – أعطيكماهن "I give them to you"
		6C	**كموهما - وناهما**	أعطيتكموهما "I gave them to you both" – أعطيتموناهما "you gave them to us both"

- **Pattern:** It is used in derivational and inflectional morphology to create new words from roots (Fabri et al. 2014). Typically, the pattern is used in inflection to combine a root with affixes, indicating the grammatical features of a stem, whereas, it is used in derivation to produce a lemma, which often leads to a change in PoS and semantic structures. For example, we can generate two lemmas: the verb "كَتَبَ" <kataba> 'write' and the noun "كَتِيبَةٌ" <katiybapN> 'troop' from the same root "كتب" <ktb> that differ from each other with respect to syntactic and semantic aspects. In these processes, the two patterns "فَعَلَ" <faEala> and "فَعِيلَةٌ" <faEiylapN> are applied, respectively.
- **Inflection:** it is the process when the root is combined with affixes (prefix, infix, and suffix) to indicate the grammatical features of a stem, such as gender, verb tense, number, and person, etc. Also, the stem may be attached with clitics such as prepositions, conjunctions. In other words, inflection guarantees that the form of the word is appropriate, and the sentence is grammatically correct without changing the meaning or part of speech of the word. Figure 3.2 presents an illustrative example of inflection paradigm by applying the pattern "مَفَاعِلُ" <mafaAEilu> on the root "درس" <drs> to generate the stem "مَدَارِسُ" <madaArisu> "schools" which is eventually attached to some clitics to have a surface word "فَمَدَارِسُهَا" <famadaArisuhaA> "and its schools".

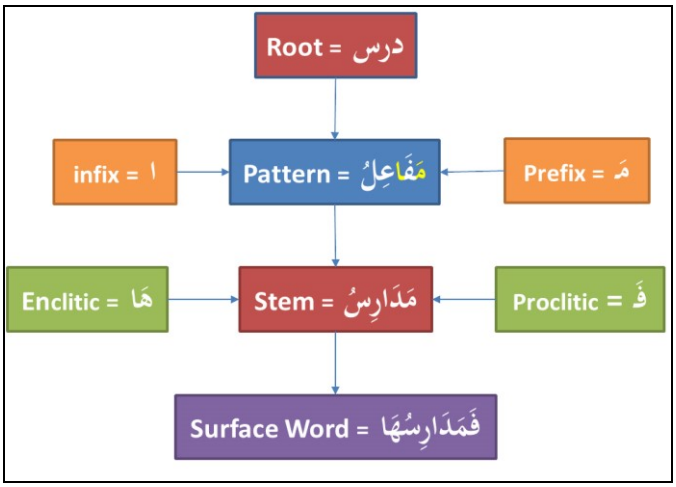

Figure 3.2 An example of an inflection paradigm

2.1.1. Classification of stemmers

Concerning the classification, there are two main types of stemmers, depending on the nature of their applied rules (Otair 2013):

1. The light stemmer is based on algorithms that remove clitics only, without trying to deal with affixes, or find roots (e.g., (Larkey et al. 2007; Zeroual and Lakhouaja 2014; Abainia et al. 2016; Aldabbas et al. 2016; Jaafar et al. 2016)).
2. The root-based/heavy stemmer uses algorithms to reduce inflected words to their roots (e.g., (Khoja and Garside 1999; Taghva et al. 2005; Al-Kabi et al. 2015)).

2.2. A Proposed Stemming System

In this section, we present a new hybrid system that takes into consideration the Arabic language features especially the interaction between the roots and patterns (rule-based), then, a statistical model is implemented to estimate the appropriate stem according to the context. As result, this system was able to overcome some issues faced by relevant Arabic stemmers.

2.2.1. Major Arabic Stemming Difficulties

Although Arabic is a Semitic language that has some specific features regarding its morphology, several Arabic stemmers are based on the same concepts used for Indo-European languages especially English. Consequently, they remove blindly the most frequent suffixes and prefixes from a surface word (e.g., (Larkey et al. 2007) and (Eldesouki et al. 2009)), which make them have a high stemming error ratio and most often result incorrect Arabic words (E. T. Al-Shammari 2013).

In English, the stem is considered as a part of the word (with or without meaning) which is used to form new words (e.g., perish-able and dur-able) where the stem may be a valid standalone word like “perish” or invalid word such as “dur” (Singh and Gupta 2016b).

In the case of Arabic, consider the verb form “يَنْتَقِلُونَ” <yanotaqiluwna> “they move out”, stemming will remove the present prefix “يَ” <ya> and the plural suffix “ونَ” <wna> and leave “نْتَقِلُ” <notaqilu> which is a non-word in Arabic. In addition, there are certain character sequences that match one of the affixes or clitics, but they are a part of the original word.

Therefore, the stemmer ends with truncating a word like "وَلَد" <walad> to "لَد" <lad> which also has no meaning in Arabic. Besides, such stemmers cannot deal with the broken plural in the Arabic language (Neme and Laporte 2013). For example, stemming the irregular noun "عُقْدَة" <Euqodap> "knot" produces the form "عقد" <Eqd> "knots" which is not its stem but its plural form.

2.2.2. Rule-based Algorithm

Historically, a root was the entry to traditional Arabic lexicons, since most Arabic words are generated from roots. The interaction between roots and patterns has intrigued lexicographers and morphologists for centuries. Unfortunately, the power of roots and patterns has not yet been fully utilized or understood in NLP (Mohammed Attia et al. 2016). Considering these characteristics of Arabic morphological structures, an algorithm is proposed in order to develop a new efficient Arabic stemmer.

Typically, a preparation of the input text is required, using the following processes:

- **Normalization**: It is used to remove punctuation characters, special characters, numbers and non-Arabic characters;
- **Tokenization**: It is used to break up the text into individual tokens using as delimiters, whitespace and newline.

The proposed stemming algorithm consists of three main phases to get all possible and correct stems and roots. These phases are:

1. **Segmentation of clitics:** This is used to remove all possible clitics from a token based on two sub-lists of clitics. Further, a table of compatibility rules between proclitics and enclitics is implemented. In doing so, we used the integrated tables in BAMA (T. Buckwalter 2004) and AL-Khalil Morpho Sys2 (Boudchiche et al. 2016) which were prepared by experts in Arabic grammar. For example, some clitics could not be attached to a word simultaneously; it means that one of the clitics or both are original characters as the case of the proclitic "سَ" and the enclitic "ي", where their combination is invalid (e.g., the word "سَاعِي" <saAEiy> 'courier'). Also, as clitics, "سَ" is for verbs, "ي" is for nouns.
2. **Checking dictionaries:** Considering that not all Arabic words came from derivational and inflectional morphology such as stop words, places names, and proper nouns, additional verification is performed in order to enhance the performance of dealing with the ambiguity. Notice that, these words do not have a root;
3. **Pattern compatibility:** In this phase, a compatibility process using patterns and roots is performed to determine the correct stems according to the Arabic language grammar. In some cases, after removing clitics, some characters attached to the end of an output return to their original shape or, more likely, the word returns to its original shape. For example:
    - "hamza" (ئ) or (ؤ) to (ء): "سماؤ+ه" <smA&+h> 'his sky' to "سماء" <smA'>;
    - ">Alif mamduwda" (ا) to (ى) "نسا+ه" <nsA+h> 'forget it' to "نسى" <nsY>;
    - "taA' mabsuwTap" (ت) to (ة) "مكتبت+ه" <mktbt+h> 'his library' to "مكتبة" <mktbp>.

Figure 3.3 presents an illustration of the rule-based algorithm.

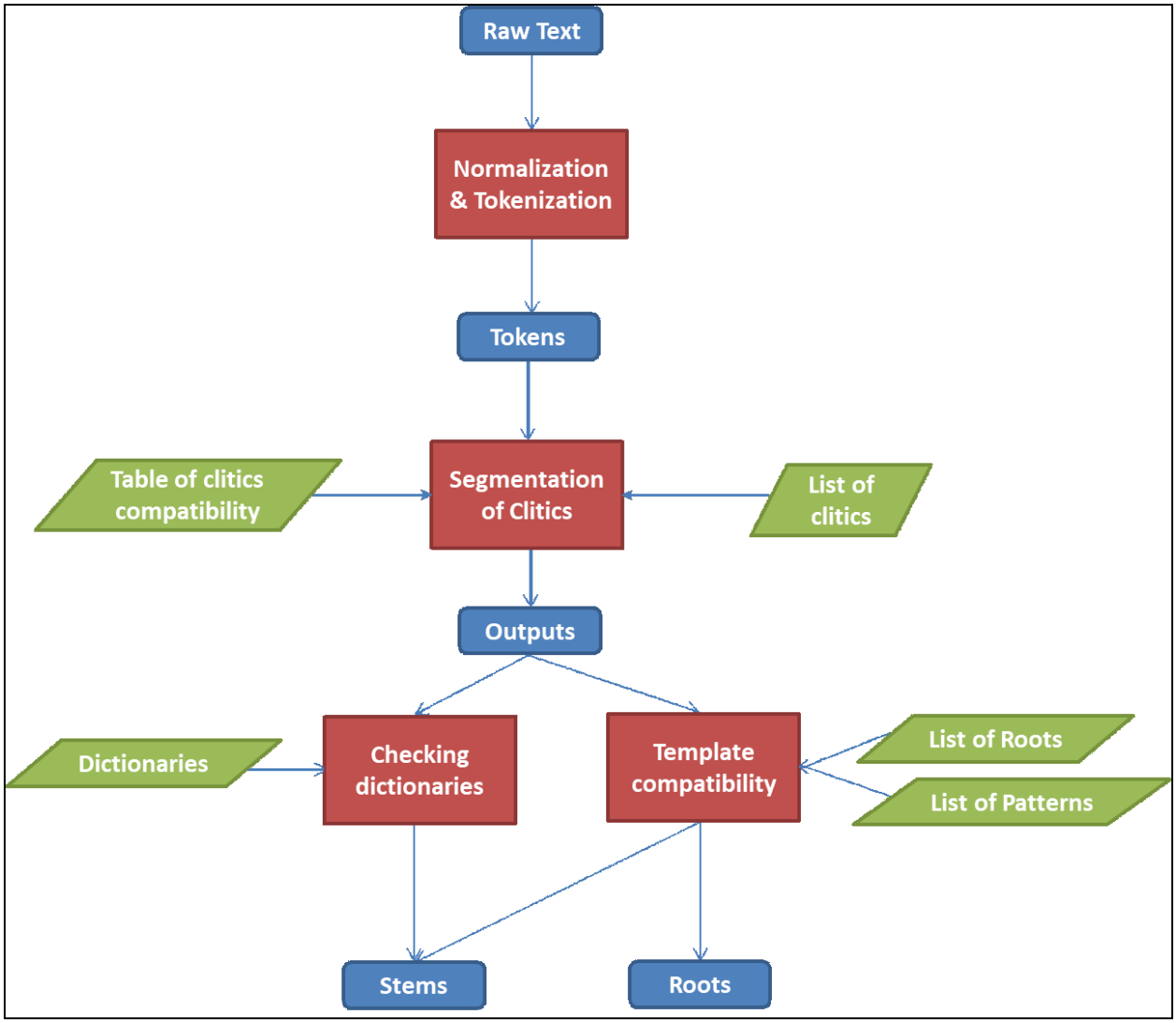

Figure 3.3 An illustrative schema of the rule-based algorithm

2.2.3. Statistical Analysis

In the previous process, the system was able to identify the potential stems for each word of a sentence. The purpose of the following statistical analysis is to select the most likely stem among these potential ones depending on the context. This process is based on the HMMs, the Viterbi algorithm, and smoothing techniques.

a) Hidden Markov Models

The Hidden Markov Models (HMMs) are used to model a phenomenon by making some assumptions about two dependent random processes. The first process consists of unobservable states (hidden states) and the second process treats the observed states. In the proposed system, the possible stems of the word obtained in the rule-based phase represent the hidden states of the HMMs model, while the unstemmed words of the sentence are the observed states of the HMMs model. The aim of HMMs is to predict the hidden states based on the observed states. For

example, the unstemmed word "كتابة" is the observed state and the hidden state is one of the possible stems such as "كِتابَة" <kitAbap> 'Writing', "كِتاب" <kitAb> 'Book', and "تَاب" <tAb> 'Repented'. Rabiner (1989) provides an overview of the basic theory of HMMs and gives more practical details.

b) Smoothing techniques

Since there is no training corpus that can cover all the transitions between Arabic words, some transition and transmission coefficients can be estimated to be zero, which is not suited to find the optimal path by the Viterbi algorithm. To overcome this issue, smoothing techniques are applied to fill these gaps and assign a non-zero probability to these coefficients of the test corpus. Based on a performance comparison of well know smoothing techniques in the literature (Chennoufi and Mazroui 2014), the Absolute Discounting method (Ney et al. 1994) is the one used in this model.

c) Viterbi algorithm

To program the Viterbi algorithm, we must first estimate the parameters of the statistical model. To do so, we applied an estimation method based on the maximum likelihood (Manning and Schütze 1999) on a tagged training corpus. Further, to find the most probable sequence of stems, we use the Viterbi algorithm (Neuhoff 1975), which is well suited for finding the optimal path. In order to get this optimal path, we make some assumptions on HMMs to attempt the maximum over all previous paths. The following Figure 3.4 exhibits an example of applying Viterbi algorithm on an Arabic sentence to find the optimal solution, which means to find the most probable sequence of stems.

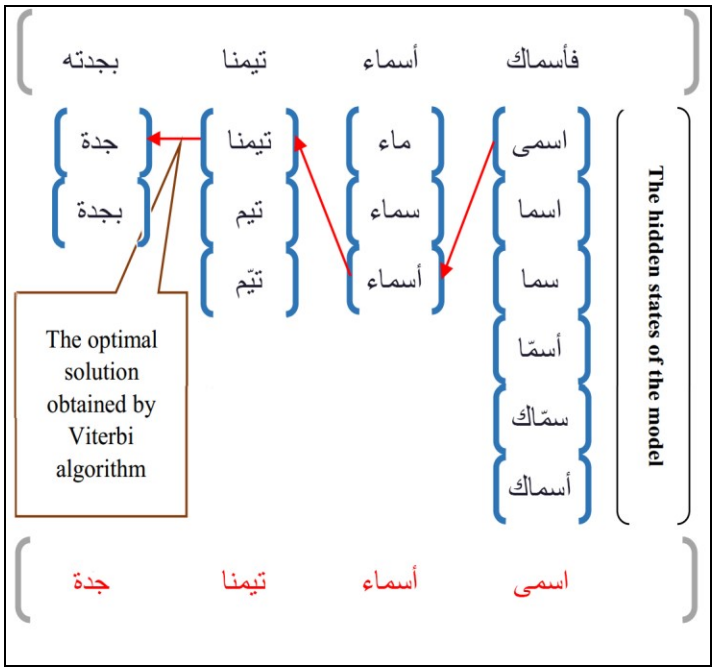

Figure 3.4 Example of applying Viterbi algorithm on an Arabic sentence

2.3. Training and Test Data

Arabic is relatively a resource-poor language when it comes to finding freely available lexical resources and pre-tagged training corpora. The following data are carefully selected to be used to train, adapt, evaluate, and compare the performance of the developed tools in this work.

The data used to build the lexicon are extracted from:

- Morphological Analysers: AlKhalil Morpho Sys2 and BAMA.
- Arabic verb conjugator: Qutrub[24];
- the Arabic Gigaword Corpus 4th Edition (Parker and Robert 2009);
- Tashkeela Corpus (Zerrouki and Balla 2017);
- Named Entities extracted from the Arabic Wikipedia[25].

The used data for the training and evaluation processes are:

- Al-Mus'haf corpus (Zeroual and Lakhouaja 2016) which represents Classical Arabic. The corpus covers the Quranic text where all the words are annotated with morphological information and PoS tagset (cf. Chapter 5, Section 2).
- NEMLAR corpus (M. Attia et al. 2005; Yaseen et al. 2006) which represents MSA. It includes texts from various domains such as Arabic literature, politics, science, sports, etc. In addition, it is divided into four parts: Raw, PoS tagged, fully vowelized, and lexically analysed.

Table 3.2 presents all these resources with more details.

Table 3.2 Linguistic resources

Type	Resources	Data description	Nb of words
Lexicon	Arabic Gigaword 4th Edition	Broken plurals	2,562
		High frequency words	37,716
	Arabic Wikipedia	Persons names	16,000
		Places names	4,587
	AlKhalil Morpho Sys	Utilities words	530
		Proper nouns	20,603
	Tashkeela	High frequency words	83,411
	Buckwalter Analyzer	Obsolete words	8,400
	Qutrub	Verbs	10,972
Corpora	Al-Mus'haf	CA corpus	78,121
	NEMLAR	MSA corpus	500,000

[24] https://qutrub.arabeyes.org/

[25] https://sourceforge.net/projects/arabicnes/

2.4. Results and Discussion

In this section, we evaluate the performance of the system we developed on both MSA and CA texts represented by Al-Mus'haf and NEMLAR corpora. Before starting this evaluation, the rule-based algorithm is applied in order to determine the average number of possible correct stems per word. To do so, a corpus of 12,345,636 words compiled from the "Islamic book library[26]". These books are intentionally chosen for two reasons, they are free and almost free of orthographic errors and have been revised which lead to more accurate results. The rate achieved is 3.51 possible stems per word. Another test is performed to know if the linguistic resources involved (cf. Section 2.3) in the rule-based algorithm truly improve its performance. For that end, a comparative evaluation is conducted with the morphological analyser BAMA (T. Buckwalter 2004) using only the NEMLAR corpus, as BAMA is addressed to MSA rather than CA texts. The aim here is to count the number of outputs given by BAMA as well as the proposed rule-based algorithm. Table 3.3 shows the obtained results.

Table 3.3 Number of obtained outputs

Stemming system	Diacritization	1 output	2 outputs	>=3 outputs
Our algorithm	Vowelled	95.1%	4.69%	0.21%
	Unvowelled	34.56%	23.89%	41.55%
BAMA	Unvowelled	45.73%	46.23%	8.04%

As noted in Table 3.3, BAMA does not deal with Arabic diacritics; therefore, it removes the diacritization marks from the vowelled words before analysing them. In the case of unvowelled words, our algorithm produces more outputs than BAMA (41.55% of words have more than 3 outputs). This is due to the richness of the integrated linguistic resources. In the case of vowelled words, our algorithm can benefit from the diacritization marks to deal properly with ambiguity. Consequently, 95.1% of the analysed words have only one output, which positively will affect further statistical analysis.

Regarding stemming in the context, our system is evaluated and compared to a morphological analyser that is well known and flexible enough to handle most ambiguous words in the Arabic language, MADAMIRA (Pasha et al. 2014b). It is worth mentioning that the training task for estimating the transition matrices and emission probabilities is performed on 90% of data, randomly chosen from the mentioned corpora. Tests are then carried out on two subsets from those corpora, and test results are shown in Table 3.4. Note that:

- the set **Rn**: consists of 10% of the remaining data from NEMLAR corpus that have not been used in the training phase.
- the set **Rq**: consists of 10% of the remaining data from Al-Mus'haf corpus that have not been used in the training phase.
- the set **Rb**: consists of both **Rn** and **Rq**.

[26] http://www.islamicbook.ws/

Table 3.4 Accuracy results

Systems	90% of training data	Subsets	Accuracy
Our system	NEMLAR	Rn	93.10%
		Rq	89.31%
		Rb	92.74%
	Al-Mus'haf	Rn	84.31%
		Rq	88.01%
		Rb	84.55%
	NEMLAR & Al-Mus'haf	Rn	**93.83%**
		Rq	**91.79%**
		Rb	**93.41%**
MADAMIRA		Rn	87.09%
		Rq	86.34%
		Rb	86.65%

As is observed in Table 3.4, we notice that:

- The best accuracies achieved are those obtained by the proposed system using both corpora for training and testing data. Consequently, our system outperforms MADAMIRA and provides better results when the training data are large and contain both text forms, CA and MSA.
- The accuracy decreased when the text form used in the training data is different from the one used in the test phase.
- At first, it was surprising that our system reached 89.31% for Rq (CA form) when the system is trained on NEMLAR (MSA form), which is better than the achieved accuracy when the system is trained on the same text form (CA). However, since the NEMLAR corpus covers over 90% of stems included in Al-Mus'haf corpus, this result is acceptable (more details in the following sections).

To sum up, the stemming system developed and presented in this section outperforms the stemming system implemented in the morphological analyser MADAMIRA.

2.5. Usability Test

In the context of usability testing, stemming is the task that impacts most directly on the performance of the part of speech tagger, primarily for morphologically complex languages. In this experiment, we evaluate the performance of a language independent tagger called Treetagger (Schmid 2013). This tagger is adaptable to any language if a lexicon and tagged training data are available. Fortunately, a recent adaptation of Treetagger for Arabic (Zeroual and Lakhouaja b

2016) is available for the public[27]. The performance of Treetagger was tested on data from the NEMLAR and Al-Mus'haf corpora. 90% of the words were used for training and the remaining 10% of the words are used for testing. The chief purpose of this experiment is to evaluate the performance of PoS tagging before and after involving the stemming process. Consequently, we have four cases:

- **Case 1:** none of the training and test data are stemmed.
- **Case 2:** training data is not stemmed, but test data is stemmed.
- **Case 3:** training data is stemmed, but the test data is not.
- **Case 4:** both training data and test data are stemmed.

The implementation of these four cases yields different results. Table 3.5 exhibits the achieved accuracy by Treetagger for each case.

Table 3.5 Improvement of accuracy results

Cases	Accuracies
Case 1	83.72%
Case 2	89.62%
Case 3	73.88%
Case 4	94.70%

Basically, the results to be compared are those obtained in Case 1 and Case 4. The accuracy achieved using only the surface form of words without stemming is 83.72%, whereas, it is 94.70% when both training and test data are stemmed. These results demonstrate that stemming process has a significant reflect on the performance of a PoS tagger. In this experiment, an improvement by 10.98% is achieved.

3. Lemmatization

Although there is confusion between stemming and lemmatization, and sometimes the terms are used interchangeably, they refer to distinct processes (Brits et al. 2005). Lemmatization is known as the process that relates words to their lemmas or a dictionary lookup form (e.g., (E. Al-Shammari and Lin 2008; Hammouda and Almarimi 2010)). The following subsection describes the lemmatization procedure in more detail and primarily for Arabic. Further, a usability test is conducted to confirm the beneficial use of lemmatization in Arabic information retrieval systems.

3.1. Lemmatization Procedure

The lemma is a lexical entry recorded in dictionaries to represent a static lexicon at a fixed point in time. Expressly, it is the state of a word when there is no conjugation "صرف" <Srf> (in case of verbs) or declension "تحويل" <tHwyl> (in case of nouns). The lemma is one specific form that represents the lexeme; the latter corresponds to a collection of all the word forms that share

[27] http://oujda-nlp-team.net/?p=2513&lang=en

the same semantic and syntactic composition. In Arabic, the lemma of verbs is chosen to be the perfective, indicative, 3rd person, masculine, and singular such as "فَازَ" <faAza> "to win". Whereas, the nominal lemma (namely, nouns and adjectives) is in the nominative, singular, and masculine (when possible), such as "لَاعِب" <laAEib> "player" and "مدرسة" <mdrasp> "School".

Regarding the Arabic language, we can say that the lemma is a result of the derivation paradigm. In contrast to stemming and inflection, the derivation is realized by applying a pattern to a root that is responsible for the choice of both syntactic and semantic structures used to produce the lemma. For instance, The following Figure 3.5 presents an illustrative example of a derivation paradigm by applying the patterns "فَعَلَ" <faEala> and "فَعِيلَة" <faEiylap> on the root "كتب" <ktb> to generate two different lemmas: the verb "كَتَبَ" <kataba> "write" and the noun "كَتِيبَة" <katiybap> "troop"; the words differ from each other with respect to syntactic and semantic aspects.

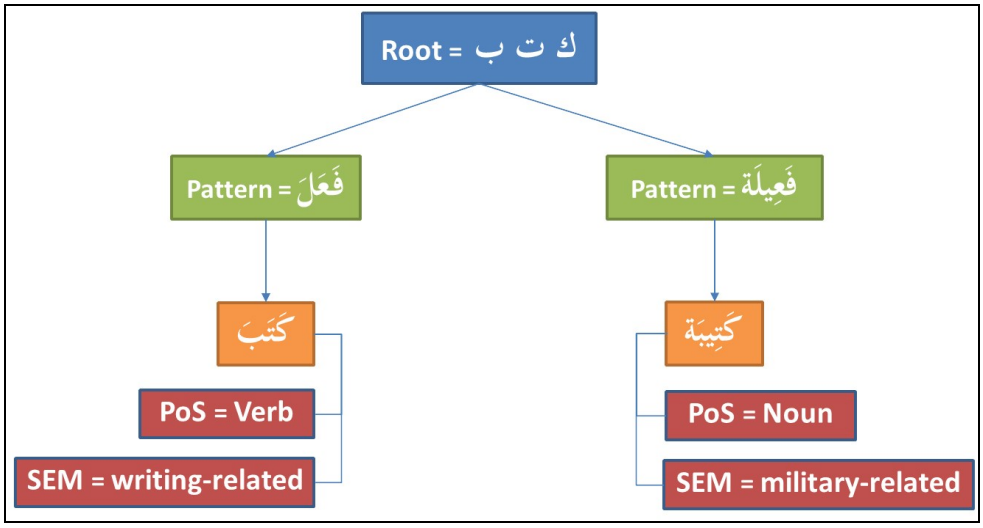

Figure 3.5 An example of a derivational paradigm

The literature in this regard is fairly limited compared to Arabic stemming. Further, unlike the stemming task, the lemmatization is a complex level of text processing. On the other hand, using lemmatization is found to be efficient, in particular, for Information Retrieval (IR), text summarization systems (El-Shishtawy and El-Ghannam 2014), text indexation (Hammouda and Almarimi 2010), and text compression (Awajan and Jrai 2015).

Based on its definition, an accurate lemmatizer requires involving large lexicon and training data as well as a statistical method. However, the reason why we did not spend time to develop our in-house lemmatizer is that, unlike stemming, there is an overall agreement on the concept of lemmatization. More frequently, a powerful lemmatizer basically requires involving large dictionaries and an effective rule-based method that considers the Arabic features such as the interaction between roots and patterns. Subsequently applying a probabilistic method like HMMs or support vector machines could lead to enhanced results. This is exactly what state-of-the-art tools do (e.g., AlKhalil lemmatizer (Boudchiche and Mazroui 2016) and MADAMIRA). Moreover, it only requires involving rich and large lexicons into some PoS taggers (e.g., Treetagger) to achieve encouraging results in terms of lemmatization.

3.2. Usability Test

In this section, we will demonstrate that both stemming (using our system) and lemmatization (using MADAMIRA) can lead to positive outcomes in Arabic IR. The present experiment is restricted to detecting stem, lemma, and surface word similarities. As a remainder, the root is the base form of both stem and lemma and an enormous number of words can be derived from it. Therefore, reducing all the surface words to a root will negatively affect the quality of obtained results (Kreaa et al. 2014). In the following experiment, the occurrences of all Quranic words, stems, and lemmas extracted from Al-Mus'haf corpus are calculated in the NEMLAR corpus, as well as in other Classical and MSA resources, namely:

- Classical resources: A collection of religion books selected from the King Abdulaziz City for Science and Technology (KACST) Arabic corpus (Al-Thubaity 2015);
- Al-Jazeera: they are articles we collected from Al-Jazeera online website[28];
- CCA: The Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell 2006). The corpus data are mainly derived from websites. The available version of this corpus includes 415 texts and covers 14 different topics.
- CNN/BBC: These both resources are a part of a large collection of corpora called OSAC, the Open Source Arabic Corpora (Saad and Ashour 2010);
- EASC: They are articles collected from newspapers, WikiNews and human-generated extractive summaries of those articles (El-Haj et al. 2015).

After stemming and lemmatization, the next step is calculating the occurrences of stem, lemma, and surface word similarities. The results of this experiment are summarized in Table 3.6.

Table 3.6 Occurrences of Quranic words in other linguistic resources

Corpora	Nb. of words	Occurrence as a surface word	Occurrence as a stem	Occurrence as a lemma
Classic	2.3 M	93.93%	99.18%	99.95%
Aljazeera	2.8 M	80.62%	95.78%	97.31%
CCA	843 K	75.76%	93.19%	95.32%
NEMLAR	500 K	75.81%	91.49%	94.65%
CNN	106 K	74.41%	91.57%	92.89%
BCC	90 K	69.62%	85.78%	88.17%
EASC	85 K	53.96%	67.50%	79.44%

In comparison to the rate obtained using the surface word, the occurrence of Quranic words is increased by an average of 14.12% and 17.46% by using respectively the stem and lemma

[28] http://www.aljazeera.net

terms. Consequently, it demonstrates the benefit of both stemming and lemmatization for Arabic IR in terms of indexing and searching tasks. In addition, the results obtained prove that the Quranic words are still utilized in MSA texts, even though there is no comparison between the stylistic level of MSA and the Holy Book.

Table 3.7 resumes the efficiency ranking of each word's form (Root, Stem, and Lemma) compared to the others in terms of storage size, grammatical features, and semantic properties.

Table 3.7 Efficiency ranking of root, stem, and lemma

	Word forms		
	Root	***Stem***	***Lemma***
Storage size	1	3	2
Grammatical features	3	1	2
Semantic and syntactic properties	3	2	1

To sum up, using the root as an indexing term will substantially optimize the storage size, but it will conflate words that differ from each other with respect to syntactic and semantic aspects. Using the stem is the best choice if we attend to conflate a set of words that share the same grammatical features. On the other hand, to unify different words that share the same syntactic and semantic properties, the lemma is the best form to choose.

4. Conclusion

In this chapter, we presented the main morphological analyses that can be applied to Arabic texts. A corpus with rich morphological analysis is very valuable for corpus linguistics and NLP, and is considered as a source of strong evidences for lexicography and linguistic description. Further, it is a preferred ground for testing the effectiveness of different IR systems. Another purpose of morphological analysis is to represent the general surface form of a word. In fact, there is no general agreement about the representation level of Arabic words. Historically, the root was the entry to traditional Arabic lexicons, since most of Arabic words are generated from roots, but many researchers have criticized this approach, and have based their representation on stems. These researchers report that precision degraded when the root is used to representing a term, due the over-semantic classification (Algarni 2016). In other words, a root conflates too many words that do not have similar semantic interpretations under one form. On the other hand, the stem form suffers from under-semantic classification. I.e., while the stem is grammatically the appropriate form regarding the context, it may exclude many similar words sharing the same semantic properties. On the contrary, the lemma refers to the set of word forms that capture the same semantic and syntactic properties, however, it may not share the same grammatical features with all these forms. To sum up, the stem is suitable for those who seek grammatical features; for syntactic and semantic properties, the lemma is the best choice.

The next chapter is dedicated to corpus annotation. There, we describe one of the essential types of annotation, the PoS tagging. Also, we provide an overview of parsing, highlighting the progress achieved in the Arabic language and mentioning relevant published works.

CHAPTER 4: Corpus Annotation

1. Introduction

One of the widely used corpus annotation forms is the Part-of-Speech (PoS) tags. Basically, it is involved in further syntactic analysis (i.e., parsing) to determine the sentence's syntactic structure. In this chapter, more details about the PoS tagging process are covered, providing numerous examples and experiments on the Arabic language. Further, relevant state-of-the-art methods and tools are presented, evaluated, and compared to our proposed methods. Finally, other annotation forms are introduced, namely the parsing and semantic analysis which are the natural successor of the PoS tagging.

2. PoS Tagging Requirements

PoS tagging is aimed at resolving ambiguity during text processing in order to assign morphosyntactic tags to each word according to the context. It is an essential task in several fields, particularly corpus linguistics and NLP. Moreover, PoS tagging reflects on the quality of other subsequent text analysis tasks primarily, parsing (Burga et al. 2013). Basically, the PoS tagging has certain requirements (Utvić 2011):

1) Selecting a suitable approach that will be used for the tagging process. Section 3 provides an overview on various methods used in this regard.
2) Defining the tagset, i.e., the basic morphosyntactic tags that will be attached to each word. In Section 4, a standard Arabic PoS tagset is proposed.
3) Preparing the required linguistic resources for training the tagger. Section 5 presents the adaptation and evaluation processes of language-independent taggers.

3. Tagging Methods

Different methods have been designed to handle the PoS ambiguity. These methods differ from each other in the approaches they are based on and the linguistic resources required. In the following subsections, we present some of the most relevant methods implemented in the PoS tagging field.

3.1. Statistical/Probabilistic Methods

In the last decades, probabilistic methods came into existence and gained more popularity because they require much less human effort. These methods are-data driven approaches based on large manually pre-tagged corpora. The training task consists of learning lexical probabilities and contextual probabilities from those corpora. A few of well-known statistical methods are listed below:

- HMM: It is an N-gram Language Model that is expressed with five parameters:
    1. the observed sequence which is the sequence of input words;
    2. the set of states (where a state is a tag);
    3. the observation sequence;

4. a matrix **A** which stores transition probabilities between states;
5. a matrix **B** which stores state observation probabilities (called emission probabilities).

Al Shamsi and Guessoum (2006) proposed an HMM PoS tagger that has achieved a state-of-the-art performance of 97%. The prediction of the tag depends only on the previous elements. However, some researchers exploit the context on both sides of a word to be tagged. For examples, the HMM tagger presented by Banko and Moore (2004), which is evaluated on both the unsupervised and supervised cases and achieves an accuracy of about 96%. Kadim and Lazrek (2016) also published a similar work. Since the purpose of their study was not to provide a new tagger to compare to existing tagging software, but rather to present a novel approach -reverse tagging- to compare and combine with direct tagging, they used only 40 sentences for the experiment phase.

- SVMs: They are support vector machines and N-gram Language Models. The SVM-based method performs a disambiguation task to estimate the appropriate solution among multiple outputs. For example, Habash and Rambow (2005) developed YAMCHA, which is an SVM-based toolkit that uses Viterbi decoding. Another SVM-based tagger called AMIRA (M. Diab 2009), which is a successor suite to the SVMTools (M. Diab et al. 2007). AMIRA employs SVMs in a sequence modelling framework using the YAMCHA toolkit. Diab reports that the tagger performs at over 96% accuracy.
- Memory-based model: an approach based on the k-nearest neighbour classifier (Cover and Hart 1967). This method performs no abstraction, which naturally allows it to deal with productive but low-frequency exceptions. It differs from most other machine learning algorithms (Daelemans et al. 1999). A tagger-generator and tagger were proposed by (Daelemans et al. 1996) based on this method and employed to produce an Arabic PoS tagger based on the Arabic Treebank corpus. It achieved an accuracy of 91.5% (Van den Bosch et al. 2007).
- Maximum entropy model (Ratnaparkhi and others 1996): one of the well-known taggers that integrates this method is the Stanford PoS tagger. This tagger achieved 96.86% of accuracy on the overall Penn Treebank and 86.91% on previously unseen words (Toutanova and Manning 2000). The last version of the tagger, which is described in (Toutanova et al. 2003) comes with trained models for other languages, including Arabic. Toutanova claims 96.42% accuracy on Arabic, training on the Arabic Penn Treebank.
- Genetic algorithm: it is a probabilistic search method that has been successfully applied in many applications of high complexity. It has been efficiently used for the solution of combinatorial optimization problems. This method is based on the principles of natural selection; which means that the search of an optimal solution is heuristic by its nature. However, compared to HMM and recurrent neural network models, the search for an optimal solution is very much longer. Ben Ali and Jarray (2013) proposed a new stochastic method based on the genetic algorithm for Arabic PoS tagging.

3.2. Rule-based Methods

The rule-based methods are developed for tagging text where the rules are hand-written by linguists (e.g., (Freeman 2001) and (AlGahtani et al. 2009)). For Arabic, a system was designed and implemented as a rule-based expert system called Qutuf (Altabba et al. 2010), and was

presented as an Arabic morphological analyser and PoS tagger. Typically, the rule-based methods are composed of three tasks:

1. Morphological analysis: it consists of an automatic stemming, tokenization, and annotation using morphosyntactic analysers. For instance, the proposed model by (Maabid et al. 2015) that uses semantic rules of the Arabic language on top of a hybrid sub-model based on two morphological analysers AL-Khalil Morpho Sys1 (Boudlal et al. 2010) and the Improved Arabic Morphology Analyser (IAMA) (Saad, E. - S.M and et al. 2005). The aim of this model is to resolve the inflected Arabic word, identify its root, find its pattern, and finally process the PoS tagging.
2. Auxiliary Lexicons: they are lexicons of words that cannot be analysed in the morphological task such as stop words, proper nouns, Arabized nouns.
3. Sentence structure: it is based mainly on the relation between the untagged words and their adjacent words (El Hadj et al. 2009). For instance, preposition and interjections are always followed by nouns. The word position in the sentence is a good indicator to identify nouns. Further, some words usually followed by nouns such as "كان وأخواتها" <kAn wAxwAthA>, "إنّ وأخواتها" <An~ wAxwAthA>, and other words mainly used with proper nouns such as "السيد" <Alsyd> "Mr.", "الجامعة" <AljAmEp> "the university", etc.

3.3. Neural Network Models

The neural network method was inspired by the Artificial Intelligence field and uses learning models inspired by the understanding of the operation of biological neural networks in brains. They typically use highly interconnected simple processing nodes (Wilson 1997). Examples of an implementation of this method in PoS tagging are (Schmid 1994a) and (Marques and Lopes 1996), and the proposed mWANN-tagger (multilingual Weightless Artificial Neural Network tagger), which is based on the WiSARD PoS-tagger (Wilkie, Stonhamand Aleksander's Recognition Device). Nevertheless, the mWANN-tagger lacked the ability to successfully tag the PoS of languages that possess nonconcatenative morphology such as Arabic which is left for a future investigation as Carneiro et al. (2015) claimed.

3.4. Hybrid Systems

Thinking that accuracy could be improved by using more than one of the previous methods, some hybrid systems were developed:

- **A combination of statistical and rule-based methods:** The objective of this combination is to employ a probabilistic model with a rule-based method or a morphological analyser in PoS tagging. For instance, the tagger developed by (Hadni et al. 2013) performed at over 97.4% accuracy using three basic tags: Noun, Verb, and Particle. Ababou and Mazroui (2016) succeeded in reaching 94% accuracy using 27 tags and providing syntactical information about proclitic attached to the words. According to Aliwy (2013), the accuracy of his statistical-based methods increased from 90.05% to 92.86% after adding a rule-based method. Aliwy claims that the low accuracy of the statistical methods is due to the small manually annotated corpus (29k words) that was implemented in his experiments.

- **A combination of a memory-based learning and rule-based methods:** the idea behind this method is to apply rules (analysing the affixes and the patterns of a word) to determine the appropriate tag of each word in the current context, then, refer to a memory-based learning method to handle the exceptions to these rules. This approach was used by Tlili-Guiassa (2006) to tag an Arabic text. In the first stage, the evaluation and all experiments of the proposed method are performed on texts extracted from educational books; later, he involved some Quranic texts which were retagged with a more detailed tagset.
- **A combination of statistical and neural network methods**: Jacob et al. (2015) have shown that the performance achieved for English using a fuzzy model over the TnT tagger is comparatively more accurate. The fuzzy model was used to overcome the performance degradation of the TnT tagger where the number of unknown words increases. To our knowledge, there is no experimental application of this method to the Arabic language.

4. Tagset

A tag is a string used as a label to describe the word's morphosyntactical features (case, gender, etc.) and a tagset is a set of these tags. The majority of tagsets used are derived from English, which is a drawback for a morphologically complex language such as Arabic. The adaptation of such tagsets is a situation for Semitic languages as Zitouni (2014) claimed: "*Approaches to PoS tagging were limited to English, resources for other languages tend to use 'tag sets', or inventories of categories that are minor modifications of the Standard English set*". Moreover, the tagsets most widely used as standard guidelines, namely those recommended by EAGLES[29], are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic. Further, several of the current systems tend to target a PoS tagset that is not sufficiently suitable for different applications (N. Habash et al. 2009) (e.g., (Khoja 2001), (Darwish 2002), and (M. Diab 2007)).

4.1. Universal Tagset

Generally, the tagsets used for each language are not identical. However, a universal tagset was proposed based on two relevant studies which cover 22 different languages including Arabic (Rambow et al. 2006; Petrov et al. 2011). Moreover, EAGLES recommendations for the morphosyntactic annotation of corpora claim that there are 13 major categories considered mandatory for most languages. After an investigation that included 30 languages from different families, we found that all these languages share 10 basic tags. Table 4.1 presents this tagset. In addition, we included "Disconnected letters", since the Arabic language is the primary focus of this research.

[29] http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#cmobli

Table 4.1 The basic tags of the universal tagset

Tags	Tag Symbols	Tags in Arabic	Examples
1. Verbs (all tenses and modes)	VERB	فعل	"كَتَبَ" (kataba "to Write")
2. Nouns	NOUN	اسم	"مَدْرَسَة" (madrasap "School")
3. Proper nouns	PN	اسم علم	"مُحَمَّد" (muHam~ad "Mohammed")
4. Pronouns	PRON	ضمير	"هِيَ" (hiya "She")
5. Adjectives	ADJ	صفة	"جَمِيل" (jamyl "Beautiful")
6. Adverbs	ADV	ظرف	"بَعْدَ، فَوْقَ" (baEda, fawoqa "After, Above")
7. Particles, Prepositions	PRT	أداة	"إلى، الذي" (<ilY, Al*y "To, who")
8. Speech-specific sounds	Uh	حرف صوت	"آه، هيهات" (|h, hayhAt)
9. Other: foreign words, abbreviations.	X	أخرى	"أوبك، مانشستر" (OPEC, Manchester)
10. Punctuation marks	SENT	علامة ترقيم	. , ;

It is worth mentioning that this universal tagset can facilitate doing several types of cross-languages studies. However, it has some limitations. It is certainly true that most English-based tagsets implemented to tag Arabic texts are not based on the linguistic reality of the Arabic grammar. Also, the universal tagset, as well as EAGLES recommendations, cannot be adapted to suit the whole Arabic syntax any better than a standard Arabic tagset. In the following, we provide some reasons that call for a standard Arabic tagset:

- In terms of morphology, due to concatenation of morphemes, where one token could represent a whole sentence through sequential concatenations, there is no "non-Arabic" tagset that could annotate such a word. For example: the token "أَسَأَلْتُمُونِيهَا" which means in English "Did you ask me for it" is normally tagged as follows:

  أَ: Interrogative common particle

  سَأَلَ: Perfect verb –active voice-

  تُ: Prominent pronoun attached to a verb

  مُ: Particle indicate plurality

  و: Particle to indicate indicative mood

  نِ: nūn of protection

  ي: Prominent object pronoun attached to a verb

  هَا: Prominent object pronoun attached to a verb

  As is seen from the example, the token has 7 tags (1 verb + 3 pronouns + 3 particles).

- The behaviour of certain categories in Arabic substantially differs from Indo-European languages and certain categories may simply not exist. For example:
    - The Gerund tag in the EAGLES recommendations is considered as a form attribute of the verb, while in Arabic it is a noun subcategory and itself has 6 subcategories (Verbal noun, Gerund with initial mïm, Gerund of instance, Gerund of state, Gerund of emphasis, Gerund of profession).
    - For the Number feature, EAGLES use only Singular and Plural. In addition to these two tags, Arabic uses "Dual" (e.g., "مدرستان" "two schools") as a tag. Moreover, the Plural in Arabic has seven subcategories (Sound plural (e.g., "عالمون" "scientists"), Broken plural (e.g., "كتب" "books"), Plural of paucity (e.g., "أشهر" "months"), Plural of multitude (e.g., "رسل" "messengers"), Ultimate plural (e.g., "قواعد" "rules"), Plural of plural (e.g., "طرقات" "roads"), and Collective noun (e.g., "قوم" "people")). Most of these types of plural have their own patterns.
    - In Arabic, the subject could be represented by its latent personal pronoun (unwritten) (ضمير مستتر/AlDamyr Almustatir).
- In terms of hierarchical level, traditional Arabic grammarians recognize only three main PoS categories (Al-Dahdah 1989) which map onto Noun "اسم" <Asm>, Verb "فعل" <fEl> and Particle "حرف" <Hrf>. Hence, all PoS tags start with those three main categories, i.e., the other EAGLES categories are subcategories of one of these three. For example, pronoun (ضمير/Dmyr), adjective (مشبهة صفة/Sft mxbht), and adverbs (ظرف/zrf) are subcategories of the noun category. Further, all the categories, not included in the noun and verb categories, are considered as particles, such as the prepositions, conjunctions, negatives, interrogatives, conditionals, etc. Besides, particles are uninflected and devoid of number, gender, and definiteness.

4.2. Arabic Standard Tagset

Unlike Indo-European languages, there is no such a standard tagset used in Arabic PoS tagging task. Consequently, it is difficult to compare and evaluate different tagging methods, especially that most of the taggers focus on their own objectives.

Several projects have been proposed an Arabic PoS tagset such as (El Hadj et al. 2009) and (Maamouri and Bies 2004). Other tagsets are derived from the Penn Treebank tagset such as (M. T. Diab 2007), and only a few works have addressed standardization. These works are (Khoja 2001), (Alqrainy 2008), (El Hadj et al. 2009), and those drawn from the Penn Arabic Treebank (PATB) (Maamouri and Bies 2004) tagset such as (Sawalha 2009) and (M. T. Diab 2007). After a comparative investigation of these tagsets, we noticed that the PATB tagset was the main ground of the other tagsets due to the following reasons:

- It is not based only on EAGLES recommendations or derived from other language such as English, rather it takes into consideration the Arabic grammar;

- The number of basic tags is 114 which makes the tagset fine-grained. It should be noted that Khoja's tagset (Khoja 2001) is based on 131 basic tags but they are derived from the BNC English tagset. Also, the Brill PoS tagger for Arabic (Freeman 2001) used a tagset of 146 tags based on the Brown corpus.

Based on the PATB tagset, Sawalha (2009) proposed a new fine-grained tagset. After that, the Qutuf team (Altabba et al. 2010) proposed a new tagset with some refinement and expansion of that of Sawalha's. Finally, this last proposal is a summary of all Arabic features, which is more theoretical than practical (Aliwy 2013). For example, tags like (صَحِيح/Sound verb) and مَضَعَّف (/Doubled verb) are difficult to determine except if we already know the morphological feature of the verb's root. Further, some tags are impossible to attribute unless if we semantically know the sentence context such as (نُونُ الْوِقايَة/nūn of protection) and (الْعَاقِل/Rational which express Humanness).

4.2.1. Criteria for a Standard Arabic Tagset

When the need for a standard tagset is invoked, one must be careful as to what criteria must be met to standardize and why. For example, one can say that some tagsets use a formal aspect while others use a more functional one. Therefore, the two should be combined into a unifying standard. Based on the findings of the previous comparative study, the proposed recommendations of EAGLES, and in collaboration with Arabic grammar experts, we propose recommendations and design criteria for morphosyntactic categories for the Arabic language considering both formal and functional aspects. These recommendations are as follows:

- Formal aspects:
    - Traditional Arabic grammar rules: The tagset should follow the Arabic grammatical system rather than those derived from other languages.
    - Identifying categories/subcategories: The ability to distinguish different levels of word categories for the morphosyntactic tagset.
    - Unambiguity: The tagset should be clearly defined.
    - Extensibility: The tagset should be easily expandable to include more Arabic features, whenever required.
    - Interchangeability: It should allow forward/backward conversion between the main categories and subcategories.
- Functional aspects:
    - Target users and/or applications: The tagset should be general enough for different applications.
    - Reusability: The tagset should be amenable to be used again by other researchers.
    - Processability: It should be possible to use a reduced version of the original tagset based on practical than theoretical reasons.
    - Comparability: It should make room for improved comparative evaluation of different PoS taggers.

4.2.2. Proposed Tagset

Based on the previous criteria, we are able to propose and evaluate a standard tagset for the Arabic language (Zeroual et al. 2017). The tagset is designed in the form of detailed hierarchical levels of categories/subcategories and their relationships. These hierarchical levels allow easier expansion when required and produce more accurate and precise results. Figures 4.1, 4.2, and 4.3 present the hierarchical levels of the noun, verb, and particle categories and their tags, respectively.

The proposed tagset has 110 basic tags classified into four distinct levels (see Table 4.2), which accurately describe and address Arabic language features considering both formal and functional aspects.

Table 4.2 Basic tags of proposed tagset

Levels	Number of basic tags			
	Noun	***Verb***	***Particle***	
Level 1	1	1	1	
Level 2	11	4	3	
Level 3	33	14	25	
Level 4	10	7	0	
Total	**55**	**26**	**29**	**110**

4.2.3. XML Structure

To make the proposed Arabic PoS tagset available and easy to use, it is encoded in XML format and made freely available to the public via our team's website[30]. Figure 4.4 displays a sample of Arabic standard PoS tagset encoded in XML format. In that figure, the main category covered is the "Noun" which is classified as a level 1 category. Then, we have some subcategories from the level 2 that share the same mode (i.e., Number) "مفرد" "Singular", "مثنى" "Dual", and "جمع" "Plural". As displayed, the plural has three subcategories from level 3 " المذكر السالم" "Masculine sound plural", "المؤنث السالم" "Feminine sound plural", and "تكسير" "Broken plural". Finally, the latter has five more subcategories namely "جموع قلة" "Plural of paucity", "جموع كثرة" "Plural of multitude", "صيغ منتهى الجموع" "Ultimate plural", "جمع الجمع" "Plural of plural", and "أسماء الجموع" "Collective noun" which are classified as level 4 categories.

[30] http://oujda-nlp-team.net/en/programms/standard-pos-tagset-arabic-language/

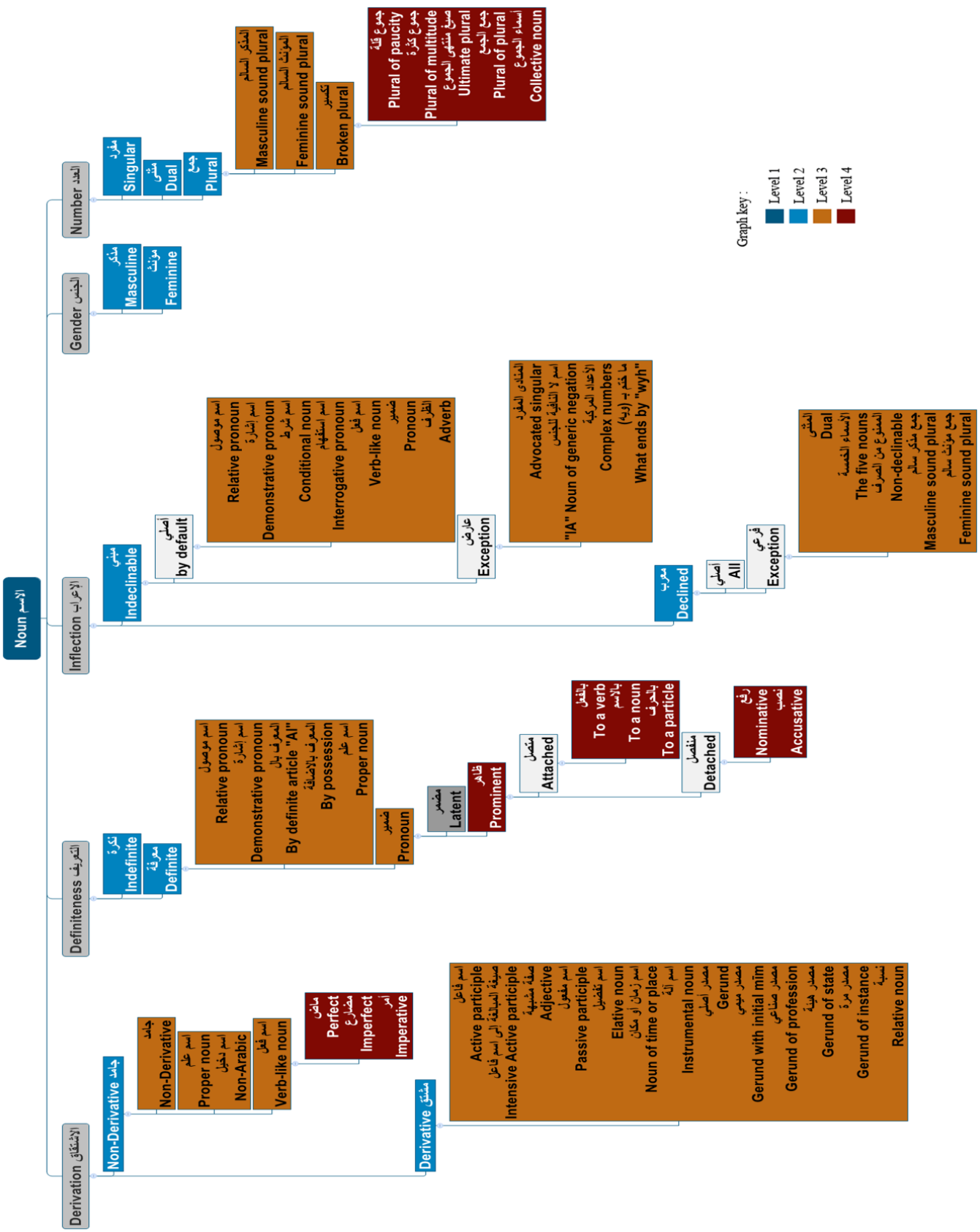

Figure 4.1 Hierarchical levels of noun categories

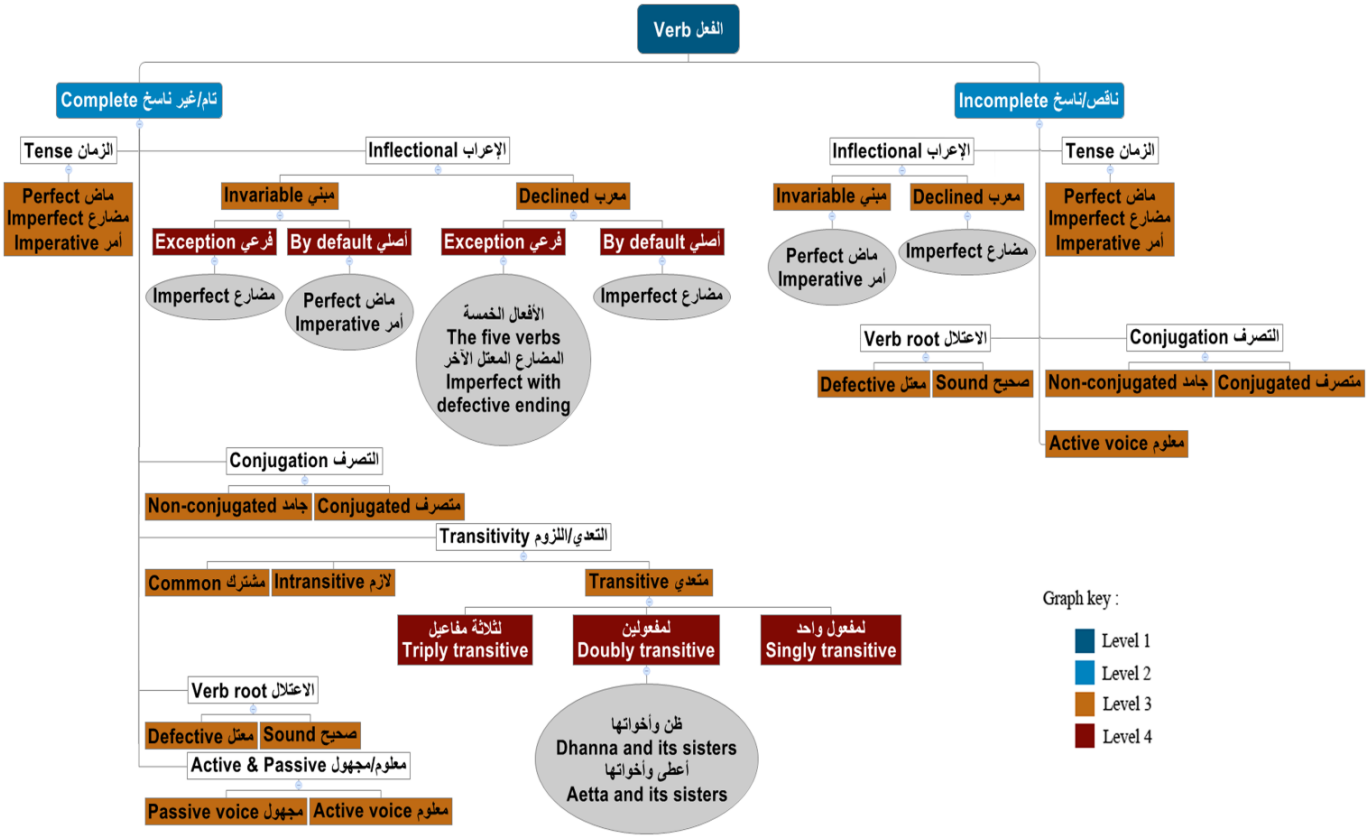

Figure 4.2 Hierarchical levels of verb categories

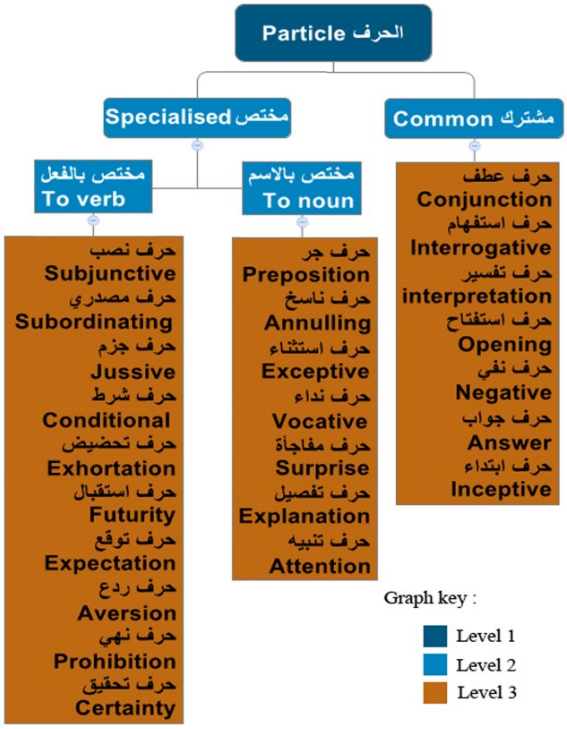

Figure 4.3 Hierarchical levels of particle categories

```
<?xml version="1.0" encoding="UTF-8" ?>
<tagset>
  <tag level='1' value_Ar='اسم' value_En='Noun'>
      <tag level='2' mod_Ar='العدد' mod_En='Number' value_Ar='مفرد' value_En='Singular'></tag>
      <tag level='2' mod_Ar='العدد' mod_En='Number' value_Ar='مثنى' value_En='Dual'></tag>
      <tag level='2' mod_Ar='العدد' mod_En='Number' value_Ar='جمع' value_En='Plural'>
          <tag level='3' value_Ar='المذكر السالم' value_En='Masculine sound plural'></tag>
          <tag level='3' value_Ar='المؤنث السالم' value_En='Feminine sound plural'></tag>
          <tag level='3' value_Ar='تكسير' value_En='Broken plural'>
              <tag level='4' value_Ar='جموع قلة' value_En='Plural of paucity'></tag>
              <tag level='4' value_Ar='جموع كثرة' value_En='Plural of multitude'></tag>
              <tag level='4' value_Ar='صيغ منتهى الجموع' value_En='Ultimate plural'></tag>
              <tag level='4' value_Ar='جمع الجمع' value_En='Plural of plural'></tag>
              <tag level='4' value_Ar='أسماء الجموع' value_En='Collective noun'></tag>
          </tag>
      </tag>
      …
  </tag>
  …
</tagset>
```

Figure 4.4 A sample of Arabic standard PoS tagset encoded in XML format

5. Language-independent Taggers

Various standard taggers have been developed based on different probabilistic methods and models and adapted to many languages. However, most of these taggers have not been officially used to PoS tagging Arabic texts. In the following subsections, we present three different standards PoS taggers:

- The TnT tagger represents the HMM.
- The Treetagger represents the implementation of a decision tree in transition probabilities to avoid problems that HMM usually face.
- The SVMTool represents the SVMs based method.

5.1. TnT Tagger

The HMM is the most widely used method for statistical PoS tagging. As a standard HMM tagger, Brants (2000) developed the TnT tagger (short form of Trigrams'n'Tags) which the transition probability depends on two preceding tags. TnT tagger uses the Viterbi algorithm for second-order Markov models. The states of the model represent tags while the outputs represent the words.

Usually, the trigram probabilities generated from training data cannot be directly used because of data sparseness. Therefore, the TnT tagger smooths the probability with linear interpolation to handle this problem. The tags of unknown words are predicted based on the word suffix. The performance of the TnT tagger was tested on two corpora, NEGRA corpus that consists of German newspaper and the Wall Street Journal (WSJ) section of the Penn Treebank corpus (Marcus et al. 1993). The reported accuracies were between 96% and 97% (Brants 2000). Practically, TnT tagger provides good efficiency when the input text consists only of known words in known context. Hence, the performance of the tagger will decline while the number of unknown words increases. During its execution, the TnT tagger runs two main programs:

- "tnt-para": is a program for the training task that requires a tagged training corpus (.tt extension) to generate the parameter file (.tnt extension). By default, it generates lexical and contextual frequencies (.lex and .123 extensions) from the training corpus.
- "tnt": is the tagging program, it requires the text file to be tagged and the lexical and contextual frequencies files (.lex and .123).

In addition to these two programs, there is an auxiliary program called "tnt-diff" for counting differences between the tagged file and a correct version of it.

5.2. Treetagger

Generally, the HMM based methods have difficulties in estimating transition probabilities accurately from limited amounts of training data. Therefore, they require a large training corpus to avoid data sparseness, and they apply different methods such as smoothing to resolve the problem of low frequencies. Consequently, by using a decision tree, a new method was developed to avoid problems that HMM face in transition probabilities. Based on this method a language independent PoS tagger called Treetagger was developed. It achieves 96.36% accuracy on Penn Treebank corpus which is better than that of a trigram tagger (96.06 %) on the same data (Schmid 1994b). The Treetagger is probably the most widely used standard PoS tagger; it has been officially and successfully used to tag about 30 different languages (Arabic not included).

Note that both tagging methods of the Treetagger and the TnT tagger calculate the probability of a tagged sequence of words recursively by:

$$p(w_1w_2\ldots w_n,\ t_1t_2\ldots t_n) := p(t_n|t_{n-2}t_{n-1})\ p(w_n|t_n)\ p(w_1w_2\ldots w_{n-1},\ t_1t_2\ldots t_{n-1})$$

Also, the Treetagger uses an unknown word PoS guesser similar to that of the TnT tagger. However, Treetagger estimates transition probabilities with a binary decision tree which mean that the probability of a given trigram is determined by following the corresponding path through the tree until a leaf is reached. Figure 4.5 shows a sample decision tree.

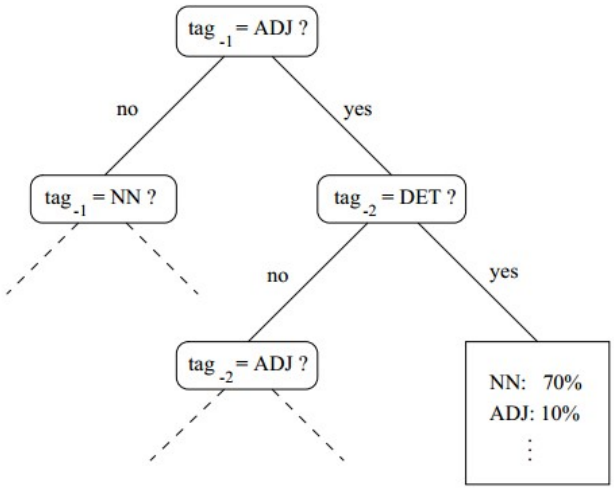

Figure 4.5 A sample decision tree

The Treetagger also runs two programs:

- "train-Treetagger" is a program for the training phase that generates the language model, i.e., a parameter file (.par), from a training corpus, a lexicon, and an open class file.
- "tree-tagging" is the tagger itself. It takes as an input the parameter file generated in the training phase and a text file to be tagged.

5.3. SVMTool

The SVMTool is proposed as a standard PoS tagger by (Giménez and Marquez 2004) based on SVMs and reported accuracy higher than all state-of-art taggers. The SVMTool comes with the implementation of five distinct kinds of models for training "0 (default),1, 2, 3 and 4" with a tagging direction that can be either "left-to-right", "right-to-left", or a combination of both. Models 0, 1, and 2 differ only in the features they consider. For example, in Model 0 the unseen context remains ambiguous unlike Model 1, which considers the unseen context already disambiguated in a previous step, while Model 2 does not consider PoS features at all for the unseen context. Model 3 and Model 4 are just like Model 0 with respect to feature extraction but examples are selected in a different manner. Model 3 is for unsupervised learning, the training is based on knowing the ambiguity class, involving a morpho-syntactic dictionary, and using PoS information only for unambiguous words. Model 4 simulates unknown words in the learning context at training time in order to learn a more realistic model.

The disambiguation complexity is controlled by introducing a lexicon extracted from the training data. Each word tag pair in the training corpus is considered as a positive case for that tag class and all other tags in the lexicon are considered negative cases for that word. This feature avoids generating useless cases for the comparison of classes. Unknown words are considered as the most ambiguous words by assigning them all open class PoS tags. The disambiguation of unknowns uses features such as prefixes, suffixes, upper case, lower case, word length, etc. Giménez and Marquez (2004) reported that the SVMTool significantly outperforms the TnT tagger under exactly the same conditions. The evaluation was for English on the Penn Treebank corpus and showed an accuracy of 97.16%. Regarding the training models,

they are based on the SVMLight implementation of the Vapnik's SVM (Vladimir and Vapnik 1995; Vapnik 2013) by Joachims (1999). Further, the tagger consists of three programs:

- "SVMTlearn" is the program responsible for the training of a set of SVM classifiers by adjusting a configuration file (config.svmt) and preparing a number of pre-tagged resources to generate the parameter files for the five training models.
- "SVMTagger" is the tagging program. It requires the path to a previously learned SVM model and a text file to be tagged.
- "SVMTeval" is the program that evaluates the performance in terms of accuracy; it needs the tagging output and the corresponding gold-standard files.

5.4. Comparative Study of Taggers

In this section, we highlight the use of the three taggers -TnT, Treetagger, and SVMTool- via various experiments and we discuss the achieved results. Regarding the tagset, we used all the four levels introduced in Subsection 4.2.2 to extend the evaluation results of the taggers. Knowing that the corpora used, Al-Mus'haf (Zeroual and Lakhouaja 2016) and NEMLAR (Yaseen et al. 2006), have different tagsets, a manual task to convert those tagsets to ours, was performed. Note that, in addition to the three main categories of level 1, we included:

- A tag for the disconnected letters that exist only in the Quranic text;
- Two tags for the punctuation signs and non-Arabic words that exist in the NEMLAR corpus.

To give an idea of the difference between the four tagsets implemented in this study, Table 4.3 presents the change of a tag from a simple level to a more fine-grained one.

Table 4.3 Illustrative examples of the implemented basic tags

Corpora	Levels	Nb. of tags	Examples
Al-Mus'haf	Level 1	4	Noun
	Level 2	26	Noun_Non-derivative
	Level 3	79	Noun_Non-derivative_Verbal
	Level 4	95	Noun_Non-derivative_Past verbal
NEMLAR	Level 1	5	Verb
	Level 2	12	Past verb
	Level 3	63	Past verb_Active voice_Transitive
	Level 4	107	Past verb_Active voice_Transitive to one object

The performance of the three taggers was tested on data from the NEMLAR and Al-Mus'haf corpora. Training data is 90% and the remaining 10% is used for testing. Table 4.4 exhibits the achieved accuracies of the taggers for each level of the tagsets.

Table 4.4 Taggers performance for each level of the tagsets

Corpora	Tested words	Unrecognized words	Levels	TnT	Treetagger	SVMTool
Al-Mus'haf	7,738	942	Level 1	95.81%	**97.18%**	94.51%
			Level 2	93.87%	**94.02%**	92.32%
			Level 3	90.82%	**91.35%**	90.45%
			Level 4	90.45%	**91.65%**	90.06%
NEMLAR	50,000	6,276	Level 1	97.16%	97.15%	**97.51%**
			Level 2	93.94%	93.86%	**95.32%**
			Level 3	92.50%	**94.74%**	94.69%
			Level 4	90.94%	**97.55%**	93.85%

The three taggers have been already implemented and evaluated for English under the same conditions. They were trained on two million words of the Wall Street Journal (WSJ) section of the Penn Treebank corpus (Marcus et al. 1993). The obtained accuracies rates are 96.06%, 96.36%, and 97.16% respectively for the TnT, Treetagger, and SVMTool. This shows that the SVMTool outperforms the other taggers. In our experiments, the obtained results show that the accuracy is influenced by the size of the training data, the tagset adopted, and unrecognized words (words are not included in training data). Furthermore, the experiments have been done on a PC dual core of 1.6 GHz with 1.5 Go RAM in Perl language and the tagging speeds achieved are shown in Table 4.5.

Table 4.5 Tagging speeds

TnT	Treetagger	SVMTool
13,700 w/s	15,000 w/s	1,000 w/s

Next, Table 4.6 summarizes the efficiency ranking of each tagger compared to the others in terms of speed, the size of the training data, and dealing with unrecognized and ambiguous words.

Table 4.6 Efficiency ranking of taggers performance

Criteria	Performance ranking		
	TnT	Treetagger	SVMTool
Training on small size of data	2	1	3
Training on medium size of data	3	1	1
Tagging unrecognized words	3	2	1
Dealing with ambiguity	2	2	1
Tagging speed	2	1	3

Generally, statistical taggers require a large training corpus to avoid data sparseness. As much as the training data is large and all senses of an ambiguous word are presented, the performance of the tagger is better. However, the tagging process achieved satisfactory results for the three taggers and for each level of the tagsets, and using a small or medium size of training data with complex tagsets did not result in a sharp degradation of the accuracy. As can be seen from the previous tables, the following points describe the advantages of each tagger compared to the others:

- Treetagger accomplished its tagging process with high speed compared to the other taggers;
- Treetagger needs less training data to achieve satisfactory accuracy as binary decision trees have relatively few parameters to estimate;
- TnT performs well on known words sequences (words included in the training set) and it gives better results than SVMTool. Still, Treetagger is better in tagging these words than both of them;
- The TnT tagger gives relatively better results than the Treetagger if they trained on medium or large data with small set of tagset, but the SVMTool does better under the same conditions;
- The SVM-based tagger outperforms the TnT tagger and Treetagger on unrecognized words (words are not included in training data), also achieved better results with ambiguity.
- Usually, the more complex tagset is used, the more the performance decreases. However, Treetagger outperforms the TnT tagger and SVM-based tagger on tagging with extensive tagset.

5.5. Accuracy Factors

During the experiments conducted, we managed to detect some factors that impact the PoS tagging accuracy. In this subsection, we highlight these factors in order to first better understand their mechanism and then to find ways to enhance the performance of the PoS tagging.

- **Tagset complexity:** Generally, the simpler the tagset is the better accuracy that will be achieved. However, based on Table 4.4, using Treetagger, the accuracy starts to increase again with more complex tagset as it is the case for Al-Mus'haf corpus (from Level 3 to Level 4) and for NEMLAR corpus (from previous levels to Level 4). As a possible explanation, the probability of a given trigram is determined by following the corresponding path through the tree until a leaf is reached. This means that if we attempt to obtain the probability of a particular tag, we must first answer the test at the root node. For this reason, a change in the tagset has a significant impact on the training process of Treetagger. For example, the probability of a tag preceded by a Verb (VERB) and a Particle (PRT) changes from one level to another. Table 4.7 presents an example of a probability change for the word "فهم" <fhm>.

Table 4.7 Example of probability change through levels

Levels	Sentences in training data	Tags for the word "فهم"	Probability
1	إِنْ جَاؤُوا فَهُمْ…	NOUN	60%
	إِنْ كَانَ فَهْمُ…	NOUN	
	إِنْ كَانَ فَهِمٌ…	NOUN	
	إِنْ اسْتَوْعَبَ فَهِمَ…	VERB	40%
	إِنْ شَاءَ فَهَّمَ…	VERB	
3	إِنْ جَاؤُوا فَهُمْ…	NOUN_Pronoun.3rd-person	20%
	إِنْ كَانَ فَهْمُ…	NOUN_Verbal-noun.Gerund	20%
	إِنْ كَانَ فَهِمٌ…	NOUN_ADJ	20%
	إِنْ اسْتَوْعَبَ فَهِمَ…	VERB_Perfect.Active-voice	40%
	إِنْ شَاءَ فَهَّمَ…	VERB_Perfect.Active-voice	

Table 4.7 presents the probabilities of a tag preceded by a verb and a particle. Even for the same word, such as "فهم" <fhm>, this can easily change based on the tag level adopted. Consequently, the change of adopted tagsets significantly affects the probabilities estimated by Treetagger during the training process, which is reflected directly in the accuracy results.

- **Tagset conversion:** Based on an investigation of the tagset effects on the PoS tagging for Arabic, Kübler and Mohamed (2012) believed that using a complex tagset and then converting the resulting annotation to a smaller tagset provides a higher accuracy than tagging using the smaller tagset directly. Fortunately, the suggested hierarchical levels also allow a similar investigation. Thus, an experiment is performed to check Kübler's investigation findings. Table 4.8 describes in detail a comparison between the achieved results using Treetagger with a smaller tagset and the achieved results after converting the complex tagset to the smaller tagset.

  All the conversion processes indicate a marginal improvement in the accuracy, supporting Kübler's investigation. More precisely, this conversion improved the performance of the PoS tagging from 0.01% (2→1 in Nemlar corpus) to 1.55% (4→1 in Al-Mus'haf corpus). This demonstrates that using complex tagsets is not necessarily an obstacle for the PoS tagging. On the contrary, they may improve the accuracy. The reason for this may be that a complex tagset precisely describes the distributional features of words. For example (see Table 4.7), the full tag (NOUN_Pronoun.3rd-person) describes the word's characteristics better than the simple tag (NOUN).

Table 4.8 Tagging accuracy with converting process

Corpora	Levels	Accuracy in direct tagging	Converting process	New accuracy
Al-Mus'haf	1	97.18%	2→1	98.20%
			3→1	98.31%
			4→1	**98.73%**
	2	94.02%	3→2	94.12%
			4→2	**94.45%**
	3	91.35%	4→3	91.69%
NEMLAR	1	97.15%	2→1	97.16%
			3→1	97.47%
			4→1	**98.03%**
	2	93.86%	3→2	94.18%
			4→2	**94.86%**
	3	95.74%	4→3	96.49%

- **Ambiguity:** The next experiment is intended to analyse the ambiguity through various levels of the tagsets. Ambiguity can exist between the main categories (Noun, Verb, and Particle) and between subcategories of the same main category. We believe that the error rate is more acceptable during the tagging process if it is between the subcategories than to be between the main categories. Table 4.9 exhibits the rate of ambiguity that is not solved during the tagging process using subcategories of level 3.

Table 4.9 Ambiguity between main categories and subcategories

Corpora	Ambiguity between main categories	Ambiguity between subcategories of level 3
Al-Mus'haf	1.27%	7.08%
NEMLAR	1.97%	3.30%

Table 4.9 emphasizes that ambiguity exists with a high degree between the subcategories (e.g. providing the adjective tag instead of the gerund tag) in comparison with the main categories of level 1 (providing the noun tag instead of the verb tag).

- **Text form:** Here we demonstrate that the text form (CA or MSA) influences the PoS taggers performance. For that end, we investigate five cases of tagger implementation using the universal tagset and new subsets of both training and testing data that differ from above datasets to extend the evaluation experiments. The five cases are as follows:

  Traditional cases:

  - Case 1: train and test the taggers on CA texts from Al-Mus'haf corpus.

- Case 2: train and test the taggers on MSA texts from NEMLAR corpus.

Unusual cases:

- Case 3: train the taggers on CA texts and test them on MSA texts.
- Case 4: train the taggers on MSA text and test them on CA texts.
- Case 5: train the taggers on a mixed training data that contain CA and MSA texts and test them on three different samples CA, MSA, and a combination of both forms.

The overall achieved accuracies are presented in Table 4.10.

Table 4.10 The influence of data training text form on tagging performance

Cases		**TnT**	**Treetagger**	**SVMTool**
Case 1: Training: Al-Mus'haf / Testing: Al-Mus'haf		93.23%	**93.32%**	92.27%
Case 2: Training: NEMLAR / Testing: NEMLAR		93.94%	93.56%	**94.88%**
Case 3: Training: NEMLAR / Testing: Al-Mus'haf		72.32%	76.83%	**78.56%**
Case 4: Training: Al-Mus'haf / Testing: NEMLAR		65.15%	**69.75%**	63.94%
Case 5: Training: Al-Mus'haf +NEMLAR	**Test 1:** Al-Mus'haf	81.99%	**82.14%**	82.06%
	Test 2: NEMLAR	91.59%	92.78%	**93.95%**
	Test 3: Al-Mus'haf + NEMLAR	87.70%	88.61%	**89.11%**

Case 1 and Case 2 show impressive accuracies, but only a 78.56% is achieved as the highest accuracy possible using SVMTool. Therefore, it is not recommended for the Arabic language to train a tagger on MSA and use it to tag a CA text or vice versa. In Case 5, the mixed training data that contain text from MSA and CA achieved better accuracies of tagging process than cases 3 and 4. However, the performance of the PoS taggers is still inferior compared to the Case 1 and Case 2. According to the previous experiments, the PoS tagger will perform better if only it is trained on the same text form. Further, to obtain high accuracy, a large training data size is required.

- **Multi-word terms:** The rate of common errors of the tree taggers varies from 2.21% to 3%. As all these taggers are developed using statistical methods, which means that the transition probability depends on preceding tags, the multi-word terms will impact the performance. For more illustration, Table 4.11 exhibits two examples of multi-word terms and their impact on the sequence of tags in the same sentence.

Table 4.11 The influence of multi-words terms on tagging performance

Sentences	**Tags order**
Mohamed First University	Noun/ Noun/ Noun/
Mohammed Ali Clay won the final	Noun/ Noun/ Noun/ Verb/ Particle/ Noun/

Since most taggers uses second-order Markov Models, the transition probability, which depends on two preceding tags, will affected in the case of multi-words terms. This kind of problem can be resolved using a combination of rule-based and statistical methods.

It is obvious that the accuracy could not reach its high-level using these adapted statistical taggers only. Further, the tagging results showed that the taggers have, to a certain degree, distinct types of features. Besides, several factors impact the performance of the tagging process. For those reasons, we highly recommend the use of a strategy that either combines the tagging results achieved by different taggers or involves other kind of tagging methods to come up with an efficient hybrid tagging system. Finally, it seems important to increase the size of the lexicon and the training corpus to decrease the number of unknown words.

5.6. Feature-rich PoS Tagging through Tagger Combination

In this section, we evaluate the performance of all possible combinations, presenting the best combination, and discuss the results achieved. Notice that, the tagset used in the following experiments is the universal tagset. In the first stage, these taggers will be evaluated individually. To do so, the taggers are trained and tested on data from the NEMLAR and Al-Mus'haf corpora. Table 4.12 exhibits the obtained accuracies from all the taggers.

Table 4.12 Accuracy results

Corpora	TnT	Treetagger	SVMTool
Al-Mus'haf	93.97%	**94.70%**	93.82%
NEMLAR	94.74%	**95.12%**	94.88%

As seen in Table 4.12, Treetagger performs better than the other taggers when they are applied on each corpus, whereas, the achieved accuracy by TnT is slightly better than the one achieved by SVMTool when it is applied on the Al-Mus'haf corpus, and vice versa when they are applied on the NEMLAR corpus. To indicate the motivation for combining taggers, a deeper investigation is required. Therefore, we checked the outputs of the three taggers to explore the common results, different errors obtained in the non-common results, and eventually to exploit these observations in further tasks. Table 4.13 shows detailed information about taggers outputs.

Table 4.13 Detailed information about taggers outputs

	Common outputs		**Non-common outputs**			
Taggers	**All**		**TnT**	**Treetagger**	**SVMTool**	**All**
Al-Mus'haf	**93.72%**		**6.28%**			
	Correct	Incorrect	Correct	Correct	Correct	Incorrect
	92.94%	0.78%	1.03%	1.76%	0.48%	3.01%
NEMLAR	**94.61%**		**5.39%**			
	Correct	Incorrect	Correct	Correct	Correct	Incorrect
	93.85%	0.76%	0.89%	1.27%	1.03%	2.20%

Several hints can be observed in this Table:

- The common outputs are not always correct; yet, the rate of incorrect ones remains very low (0.76%-0.78%).
- None of the common and correct outputs (93.72% and 94.61%) reach the accuracy rate of the three taggers individually.
- The non-common outputs (6.28% and 5.39%) are not always incorrect, only 3.01% and 2.20% respectively. I.e., more than half of them are correct.

Based on these observations, we deduce that depending only on the common outputs is not effective, because it does not reach the performance level of each tagger individually. Also, we cannot abandon the non-common outputs, where there are an interesting percentage of correct results. Therefore, it is possible to define an appropriate combination algorithm. Thus, the purpose of this work is to propose a combination algorithm, and to verify if it does effectively improve tagging accuracy. Here, we describe the algorithm implemented for the combination process. This combination algorithm determines the most appropriate tags in three steps:

1. Tagging the input text with all taggers;
2. Selecting for each token the most voted tag from the majority taggers (in these experiments, at least two taggers);
3. If the given tags from all taggers are unlike. Then, the selected tag is the one proposed by the most accurate tagger (in these experiments, is Treetagger).

The evaluation of the algorithm is divided into two phases. In the first one, only two taggers are used in the combination; consequently, we are left with three possible combinations. In the second phase, the three taggers are used as a combination. Table 4.14 shows the achieved accuracies of all combinations in these two phases.

Table 4.14 Combination accuracies

Combinations	TnT & Treetagger	Treetagger & SVMTool	TnT & SVMTool	All taggers
Al-Mus'haf	95.73% (+)	93.54% (-)	93.82% (-)	**95.79% (*)**
NEMLAR	95.23% (+)	95.00% (-)	94.93% (+)	**96.45% (*)**

By combining the outputs of two or three taggers using the proposed algorithm, the results obtained are as follows:

- (-): this combination achieves an accuracy rate lower than the most accurate tagger involved in this combination.
- (+): this combination achieves an accuracy rate higher than the most accurate tagger involved in this combination.
- (*): the best achieved result in all combinations; i.e., those involve all the three taggers.

After testing and validating the combination algorithm on pre-tagged corpora, what remains is to evaluate it on new untagged/unseen data which is the main objective of this work. For that reason, we have selected the data from a resource that is rich in terms of a variety of

domains and topics. The data are extracted from the Arabic part of the MulTed corpus (cf. Chapter 5, Section 4), a new proposed multilingual corpus constructed based on the available subtitles of TED talks. The size of these data is 500,000 tokens.

Before applying the combination algorithm, it is required to determine the most accurate tagger among the three. This is based on the idea that the tagger which outperforms the others in the non-common outputs will be the one that has the highest accuracy in the overall corpus. Hence, the first stage of this evaluation is to annotate the corpus with the three taggers and to separate the common and non-common outputs. Finally, we manually verify and validate the achieved accuracies in two experimental samples:

1) all non-common outputs;
2) 10% of random common outputs.

Table 4.15 presents the obtained results of this task.

Table 4.15 Accuracy analysis on experimental samples

	Common outputs		**Non-common outputs**			
Taggers	*All*		*TnT*	*Treetagger*	*SVMTool*	*All*
Percentages	86.98%		13.02%			
Experimental samples	10.00%		13.02%			
Correctness	*Correct*	*Incorrect*	*Correct*	*Correct*	*Correct*	*Incorrect*
Accuracy	84.85%	2.13%	**4.03%**	4.02%	3.54%	1.43%

The hints observed in the previous evaluations (Table 4.13) remain the same as they are in the experimental samples presented in Table 4.15 except that the TnT slightly outperforms the other taggers on non-common outputs. Therefore, the next step is to apply the combination algorithm and compare it to the performance of each tagger. Since the verification is done manually, only both experimental samples (23.02%) are used instead of the overall corpus. Table 4.16 exhibits the obtained accuracy rates by all the taggers and the combination algorithm.

Table 4.16 Taggers accuracies on the Arabic part of MulTed corpus

Taggers	**TnT**	**Treetagger**	**SVMTool**	**Combination**
Accuracy	88.88%	88.87%	88.39%	**90.63%**

As seen in Table 4.16, the accuracies achieved by the three taggers are approximately similar, with better performance being obtained using the TnT tagger. However, the combination algorithm outperforms the three taggers individually. Consequently, the combination system does effectively improve tagging accuracy considering the number of taggers involved and their performance. To sum up, the most important results obtained in this investigation, we state the following points:

- As seen in the evaluation experiments, the proposed combination system performs better than the other taggers when they applied individually on all three corpora.

- Usually, the PoS tagging is done by an automatic process and manually corrected afterward. To minimize the hand-correction, the combination algorithm can be used to improve the accuracy rate, and to point the candidate mis-tagged words by indicating the unlike tag predictions.
- By combining only two taggers, an accuracy rate reduction could be achieved, yet, in our case, the accuracy was lower than the most accurate tagger involved in the combination algorithm.
- Improving the performance of the current combination algorithm is at hand. For instance, the improvement still possible if the number of involved taggers is augmented or different combination algorithms are adopted.

In addition, all results obtained show that the accuracy of a common output is always lower than that achieved by the taggers separately. The reason is that the taggers produce different errors and these differences are exploited in the combination to yield better results. Therefore, we suggest combining PoS taggers, especially those produce different kind of errors.

6. Other Annotation Forms

Alongside PoS tags, there are other important annotations forms namely parsing and semantic analyses, which normally are the natural successor of PoS tagging. However, they require a good understanding of grammar rules and semantic knowledge. Due to the particularity of the Arabic grammar and the lack of free parsed or semantically annotated corpora, the Arabic language has so far received little research on the level of parsing and semantic analyses.

6.1. Parsing

Parsing, primarily dependency parsing, is the natural successor to PoS tagging. Basically, it provides a dependency tree as an output. The goal of parsing is to predict for each sentence or clause its syntactic structure. The latter is an abstract representation of the grammatical entities and liaisons between a sentence's words. Consequently, the aim of a parser is to assign a fully labelled syntactic tree to sentences of a corpus (Tsarfaty et al. 2013). For example, Figure 4. 6 exhibits a parse tree of the Arabic sentence "ذَهَبَ الوَلَدُ إِلَى مَدْرَسَةِ الحَي" "The boy went to the district's school".

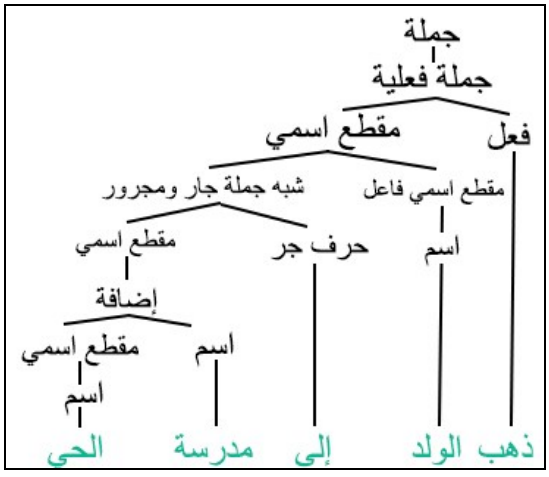

Figure 4. 6 A parse tree of Arabic sentence

Basically, parsing is a process that requires innovative methods to deal with several contexts and take into consideration a language's grammar and its lexical features. Therefore, existing Arabic parsers had a week standing, especially since Arabic is very rich in terms of morphological complexity and grammatical features. However, recent promising works have been published and present new methods to deal with Arabic texts. The best examples to give in this context are those proposed by Manning et al. (2014), Dukes (2015), and Ababou (2017), which use machine learning for statistical parsing for the Arabic language.

6.2. Semantic Analysis

Due to the tremendous growth in Arabic content on the web especially in news websites and social media, both semantic and sentiment analyses have received a great deal of attention from several scientific groups to collect huge amount of data as input for their studies over the last few years. Furthermore, promising semantic Arabic lexicons have been released, like Arabic WordNet (Regragui et al. 2016) and Azhary (Ishkewy et al. 2014). These lexicons group Arabic words into synsets and record several semantic relations between these words such as synonymy, meronymy and antonymy. Now researchers can depend on such lexicons for the automated construction of a lexical ontology for the Arabic language.

In 2001, semantic analysis was used for the first time as a conceptual model for the web content, and led to the birth of a new trend of research under the name of semantic Web (Berners-Lee et al. 2001). The aim here is making the web contents be readable and more understandable for machines, especially web crawlers and search engines. Basically, integrating semantic annotation, as well as other language data, into the Web will extend and improve NLP, computational linguistics, IR, and question answering systems. Further, this kind of annotation form allows adding information that link words with relationship values, especially if a specific topic is targeted. Semantic annotation tools provide different means to represent the content for machine processing. To that end, the Web Ontology Language (OWL) is used as one of the most common formats for ontological representation of concepts, their relationship, and semantic rules that could be applied to the knowledge. For instance, the following Figure 4.7 provides a part of ontology for food and recipe.

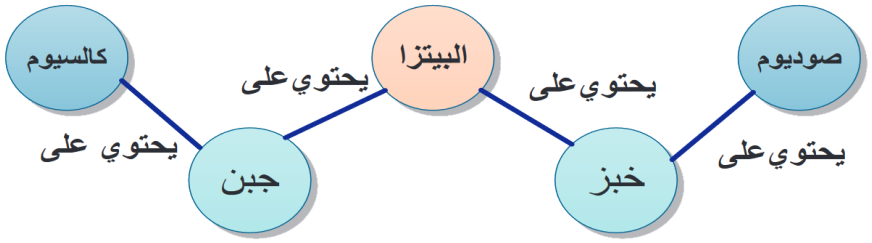

Figure 4.7 Part of ontology for food and recipe (Al-Bukhitan et al. 2014)

This figure shows some words that represent food and their recipes. Expressly, "البيتزا" <AlbytzA> "Pizza" includes "خبز" <xbz> "bread" and "جبن" <jbn> "cheese"; similarly, these letters include "صوديوم" <Swdywm> "sodium" and "كالسيوم" <kAlsywm> "calcium" respectively.

7. Conclusion

In this chapter, we described one of the essential types of corpus annotation: PoS tagging. Many approaches to perform the PoS tagging task and various relevant Arabic PoS taggers were presented. Furthermore, we discussed the well-known tagsets used in this field and their drawbacks. Subsequently, a range of criteria were listed that are useful in building a standard tagset for general use and is suitable for both forms of Arabic: Classical and Modern Standard. A usability test was performed and involved adapting three relevant language-independent PoS taggers, TnT, Treetagger, and SVMTool. The main purpose was to evaluate these taggers and apply them to the Arabic language. Regarding the comparative study, many factors influenced the performance of each tagger compared to the others such as, the size of training data, the complexity of the tagset implemented, and the text form. For instance, the Treetagger outperformed the other taggers when a small training corpus and a detailed tagset were used. On the other hand, SVMTool did better with a large training corpus and a small tagset. It was obvious that accuracy could not reach its highest-level using any of these adapted statistical taggers individually. In addition, tagging results showed that the taggers have, to a certain degree, distinct types of behaviours. For those reasons, we suggested and tested a strategy that combines the three taggers and achieved somewhat better results. It is worth mentioning, the results achieved using those adapted PoS taggers are satisfactory and respond to our critics. Thus, unlike stemming, the need to develop a new tool was not great.

Finally, this chapter also introduced other annotation forms, namely parsing and semantic analysis, highlighting the progress achieved in the Arabic language and mentioning relevant published works. Next, we describe the corpora developed during this thesis and their contribution to the literature.

CHAPTER 5: Developed Corpora

1. Introduction

From the start we were keen to build distinct types of corpora, annotated and produced in most suitable machine-readable forms. To that end, we have included both written and transcribed materials, relying on different CA and MSA resources and covering diverse topics. This chapter is devoted to presenting the three corpora developed during this thesis, describing the procedures used in their building and characteristics. Finally, the resulting corpora are compared to similar state-of-the-art corpora, stressing the significant of their contribution to the literature.

2. Al Mus'haf Corpus

The choice of Quranic text, as a starting point for working in Arabic NLP, was made carefully. The Quran, the Holy Book of Islam, had a significant impact on the Arabic language. Watson (2002) has shown that the Classical Arabic was based primarily on the language of the western Hijazi tribe of Quraysh, with some interference from pre-Islamic poetic koiné (Ferguson 1959) and eastern dialects. However, with the rise of the new religion of Islam, the Arabic language was codified in the Quran. Further, the Quranic text is a part of almost all Arabic corpora. As illustrated in Chapter 3, morphological processing of Arabic text is often handicapped by subtle orthographic issues. Unfortunately, almost all the contemporary texts are written without the diacritics. Consequently, the same word may be spelled in different ways, especially when the various dialects are considered, in which standardized forms do not exist (N. Habash et al. 2012). Thus, another reason for working with the Quranic text is the full diacritical marks it contains, which make it easy to have a precise phonetic representation of Arabic. Also, the Quranic text is the most formal and standard form of Arabic.

Given the importance of the Quranic text for the Arabic language, several researchers have been interested in building a Quranic corpus. Among the developed Quranic corpora (Section 5.2.1, Chapter 2), there are two that share with us the purpose of building a morphosyntactically annotated Quranic corpus:

- "Quran Corpus of Haifa" an offline morphological analysis performed at the University of Haifa (Dror et al. 2004).
- "Quranic Arabic Corpus", the result of an online-annotated linguistic resource from the University of Leeds (Dukes and Habash 2010).

In this section, we present the Al Mus'haf corpus (Zeroual and Lakhouajaa 2016). The corpus is annotated using rich morphosyntactical information such as stems, PoS tags, lemmas, roots, and the vowelled patterns for each of the stems and lemmas. The current version of the Al Mus'haf corpus is released under the Creative Commons Attribution 4.0 License (CC-BY-ND)[31] with the International Standard Language Resource Number (ISLRN) 114-868-598-820-5. It is free for download on the following site: http://oujda-nlp-team.net/en/programms/al-mushaf-corpus/.

[31] https://creativecommons.org/licenses/by-nd/4.0/

2.1. Methodology

To build a new corpus of the Quran, we used the Quranic text written in Uthmaani in Hafs script as it is the most common version in the Islamic world today. The annotation process has been done using the second version of "AlKhalil Morpho Sys2" (Boudchiche et al. 2016) followed by a manual verification under the supervision of expert linguists. Unlike any other Arabic texts, the Quran contains the marks of pause and intonation. Usually, the shape of some letters changes when it is concatenated with such marks. Therefore, all these marks are removed, for two reasons:

- As all NLP tools, AlKhalil Morpho Sys cannot analyse words attached to these marks;
- The isolated Quranic words are often pronounced without the use of the punctuation and intonation symbols.

The other important thing in this regard is that the letter "alef" "ا" may be placed over some letters using the "dagger alef" " ٰ " <`> U+0670, like in the word "صَلَوٰت" <Salawa`t> "prayers" "هَٰذَا" <ha*aA> "this", "أُولَٰئِكَ" <Auwla`}ika> "those" and "الرَّحۡمَٰنِ" <Alr~aHoma`ni> "the entirely merciful". To be able to analyse these words, we have modified AlKhalil to consider this orthography.

Basically, AlKhalil analyses the input text out of context, consequently, it produces multiple outputs if analysing unvowelled words. However, if the input word is vowelled, as the Quranic words, the number of outputs is reduced. In fact, AlKhalil has analysed 94% of the Quranic words at a rate of 1.65 outputs per each input word. Due to the complexity of Arabic morphology and the special properties of Quranic texts, a manual treatment for identifying and annotating the words considering context was required. This led us to manually cheek the results and obtain the following:

- 6% of words are not analysed.
- 16.3% of input words have multiple outputs and contain the correct analysis according to the context;
- 5.7% of input words have multiple outputs but none is the correct analysis according to the context;
- 3.75% of inputs words have one output that is not the correct analysis according to the context.

Table 5.1 shows more details about these four different cases.

Table 5.1 Statistics about AlKhalil analysis of the Quranic text

Distinct Quranic words	Outputs		Percentage
17,455 (100%)	Non-analysed words		6%
	One output (72%)	Correct	97.3%
		Wrong	2.7%
	multiple analyses (22%)	2 outputs	44.5%
		3 outputs	22%
		4 outputs	11%
		5 outputs	2%
		6 outputs	2.5%
		7 outputs	0.7%
		>=8 outputs	17.3%

2.2. Comparative Study

The most important thing that distinguishes the Al Mus'haf corpus from other corpora is that all the words are annotated with rich morphosyntactic information. Table 5.2 presents a comparison in terms of morphosyntactic information between the three corpora: Al Mus'haf, "Quranic Arabic Corpus" (Leeds), and "Quran Corpus" (Haifa).

Table 5.2 A comparison of morphosyntactic information

Corpora / Morphosyntactic information		Haifa	Leeds	Al Mus'haf
Number of roots		1000	1644	1673
Number of stem patterns		100	#	1357
Number of lemma patterns		#	#	244
PoS tags number	Nominal	13	12	57
	Verbal	6	6	11
	Particles	8	34	43

As shown in the previous table, the Al Mus'haf corpus is characterized by notable features which distinguish our corpus from other corpora. In comparison to these corpora, the automatic processing of "Quran Corpus of Haifa" is not complete. It remains manually unverified and has multiple outputs for each word in the final published data-set. Based on considering a random

sample, the authors estimate the final accuracy of annotation to have an F-measure of 86%. On the other hand, the "Quranic Arabic Corpus" is manually verified and computationally analysed but it still has several problems. Next, we list a number of impurities that appear in both corpora and have been addressed in the Al Mus'haf corpus.

- For the roots:
    - We have managed to determine a number of roots superior to the number of roots in the other Quranic corpora;
    - Approximately 40% of Arabic roots are missing in the Haifa corpus;
    - Several mistakes have been observed in some roots. For instance, for the words "تَطْمَئِنُّ" <taToma}in~u> "find rest", "صَوامِع" <SawAmiE> "minarets", and "سُلْطانا" <suloTAnA> "Authority" the given roots are, respectively, "طمن" <Tmn>, "صمع" <SmE>, and "سلط" <slT> instead of the correct ones "طمءن" <Tm'n>, "صومع" <SwmE>, and "سلطن" <slTn>.
- For the lemmas:
    - Lemmas are missing in the Haifa corpus;
    - The majority of lemmas in the Leeds corpus are in fact stems. For example, the words "عَالَمِينَ" <EAlamiyna> "Worlds", "ظُلُمَات" <zulumAt> "Darkness", and "كافِرونَ" <kAfirwna> "Disbelievers", are described in the corpus as lemmas, while they are inflected words. The correct lemmas are, respectively, "عالَم" <EAlam> "World", "ظُلْمَة" <zulomap> "Darkness", and "كافِر" <kAfir> "Disbeliever".
- For the patterns:
    - Stem patterns and lemma patterns are not provided in the Leeds corpus;
    - The Haifa corpus contains only a few stem patterns and they are not vowelled.
- For the PoS tagset:
    - For both corpora, the tagset used in annotation is not rich enough to represent the grammatical features of Arabic language. For example, some PoS tags are missing like "المصدر" "Gerund/Verbal noun", "اسم الفاعل" "Active participle", "اسم التفضيل" "Elative noun", and "اسم الآلة" "Instrumental noun". On the other hand, Al Mus'haf corpus is annotated using a standard tagset whose tags are carefully selected (Section 4.2, Chapter 4).

2.3. Corpus Format

To make the Al Mus'haf corpus easy to use, we produced three different formats Text, CSV and XML. Each word is associated with its morphosyntactic information in addition to the verse number and the chapter name, local, and order. We also use some symbols to facilitate reading the corpus such as:

- The '/' symbol to separate between the Arabic script and its Buckwalter transliteration. Also, between the Arabic PoS tag and its English translation;
- The '#' symbol for the non-existant information (for example, the particles do not have roots or patterns);
- The '|' symbol to separate between the morphosyntactic information in the raw text format.

Figures 1 and 2 in Appendix "A" display samples of the Al Mus'haf corpus in TXT and CSV formats. Figure 5.1 below displays a sample encoded in XML format. It exhibits the second verse "الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ" "[All] praise is [due] to Allah, Lord of the worlds" of the first chapter in the Quran "الْفَاتِحَة" "Al-Fatiha".

```
<?xml version="1.0" encoding="UTF-8" ?>
<Quran>
  <Chapter Name="سُورَةُ الْفَاتِحَةُ" Number="1" Local="مكيّة">
  …
        <Word Verse_number="2" is_basmalla="0" Vowlledform="الْحَمْدُ /AloHamodu" Stem="حمد/Hmd"
        StemPattern="فَعْلُ/faEolu" POS="مأ/VN" Lemma="حَمْد/Hamod" LemmaPattern="فَعْل/faEol"
        Root="حمد/Hmd" />
        <Word Vowlledform="لِلَّهِ/lil~ahi" Stem="لله/llh" StemPattern="#" POS="اسجل/MN" Lemma="لله/llh"
        LemmaPattern="#" Root="#" />
        <Word Verse_number="2" is_basmalla="0" Vowlledform="رَبِّ/rab~i" Stem="رب/rb"
        StemPattern="فَعْلِ/faEoli" POS="اسج/NN" Lemma="رَبّ/~rab" LemmaPattern="فَعْل/faEol" Root="ربب/rbb"
        />
        <Word Verse_number="2" is_basmalla="0" Vowlledform="الْعَالَمِينَ/AloEaAlamiyna" Stem="عالمين/Ealmyn"
        StemPattern="فَاعَلِينَ/faAEaliyna" POS="اسج/NN" Lemma="عَالَم/EaAlam" LemmaPattern="فَاعَل/faAEal"
        Root="علم/Elm" />
  …
  </Chapter>
…
</Quran>
```

Figure 5.1 A sample of Al Mus'haf corpus encoded in XML format

Moreover, a multilanguage separate XML file is generated which contains all Quranic verses and their translations to various languages like English, French, and Spanish. Further, metadata is included in this file to provide information about Chapters and verses. These

translations are extracted from the Tanzil project[32] as well as the original Arabic Quranic text. Figure 5.2 gives a sample of this file.

```
<?xml version="1.0" encoding="UTF-8" ?>
<Quran>
  <Chapter Name="سُورَةُ الْفَاتِحَةُ" Number="1" Local="مكيّة">
        <Verse Number="1" textAr="بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ" textEn="In the name of Allah, Most Gracious, Most
        Merciful." textFr="Au nom d'Allah, le Tout Miséricordieux, le Très Miséricordieux." textEs="En el
        nombre de Dios, el Compasivo con toda la creación, el Misericordioso con los creyentes." />
        …
  </Chapter>
…
</Quran>
```

Figure 5.2 A sample of the Al Mus'haf corpus translations encoded in XML format

3. OSIAN Corpus

The World Wide Web has become a fundamental resource for building large text corpora. Broadcasting platforms such as news websites are rich sources of data regarding diverse topics and form a valuable foundation for research. Although the Arabic language is extensively utilized on the Web, it still is an under-resourced language in terms of availability of freely annotated corpora. This paves the way for us to build a large corpus based on international broadcasting platforms content. This section presents the first version of the **O**pen **S**ource **I**nternational **A**rabic **N**ews (OSIAN) corpus. The corpus data was collected from international Arabic news websites and consists of about 3.5 million articles comprising more than 37 million sentences and roughly 1 billion tokens. The corpus is encoded in XML where each article is annotated with metadata information. Moreover, each word is annotated with lemma and part-of-speech.

The prime motivation for building the OSIAN corpus comes from the lack of open-source Arabic corpora that can cope with the perspectives of Arabic NLP and IR, among other research areas. Yet, we expect that the OSIAN corpus can be used to answer relevant research questions in corpus linguistics, especially investigating variation and distinction between international and national news broadcasting platforms with a diachronic and geographical perspective.

3.1. Literature Review

For almost a decade, the World Wide Web has become increasingly a source for researchers, in particular individuals, interested in the compilation of very large web-derived corpora in a short time and with low cost (Nakov 2014). As seen previously, our survey (Zeroual and Lakhouaja 2018) reports that 51% of corpora are constructed based, totally or partially, on the web content.

The web corpora continue to gain relevance within computational and theoretical linguistics. Given their size and the variety of domains covered, using web-derived corpora is

[32] http://tanzil.net/trans/

another way to overcome typical problems faced by statistical corpus-based studies such as data-sparseness and lack of variation. Moreover, they can be used to evaluate different approaches for the classification of web documents and content by text genre and topic area (e.g., (Chouigui et al. 2017)). Web corpora have also become a prime and well-established source for lexicographers to create many large and various dictionaries using specialised tools such as the corpus query and corpus management tool Sketch-Engine (Kovář et al. 2016). Moreover, some completely new areas of research, which deal exclusively with web corpora, have emerged. Indeed, the aim was to build, investigate, and analyse corpora based on online social networks posts, short messages, and online forum discussions.

Publicly available Arabic web corpora are quite limited, which greatly impacts research and development of Arabic NLP and IR. However, some research groups have shown potential in building web-derived corpora in recent years. To name a few:

- Open Source Arabic Corpora[33] (**OSAC**) (Saad and Ashour 2010): It is a collection of large and free accessible raw corpora. The OSAC corpus consists of web documents extracted from over 25 Arabic websites using the open source offline explorer, *HTTrack*. The compilation procedure involves converting HTML/XML files into UTF-8 encoding using "Text Encoding Converter" as well as removing the HTML/XML tags. The final version of the corpus comprises roughly 113 million tokens. Besides, it covers several topics namely Economy, History, Education, Religion, Sport, Health, Astronomy, Law, Stories, and Cooking Recipes.
- **arTenTen** (Arts et al. 2014): It is a member of the TenTen Corpus Family (Jakubíček et al. 2013). The arTenTen is a web-derived corpus of Arabic crawled using Spiderling (Suchomel et al. 2012). The arTenTen corpus is partially tagged. I.e., one sample of the corpus, comprising roughly 30 million, is PoS tagged using the Stanford Arabic part-of-speech tagger, and another sample, containing over 115 million words, is tokenised, lemmatised, and PoS tagged using the MADA system. All in all, the arTenTen comprises 5.8 billion words but it can only be explored by paying a fee via the Sketch Engine website[34].
- **ArabicWeb16**: Since 2009, the ClueWeb09 (Callan et al. 2009), that includes 29.2 million of Arabic pages, was considered the only and largest Arabic web crawl available. However, in 2016, a new and larger crawl of today's Arabic web is publicly available. This web crawl is called ArabicWeb16 (Swuaileh et al. 2016) and comprises over 150M web pages crawled over the month of January 2016. In addition to addressing the limitation of the ClueWeb09, ArabicWeb16 covers both dialectal and Modern Standard Arabic. Finally, the total size of the compressed dataset of ArabicWeb16 is about 2TB and it is available for download after filling a request form[35].
- The **GDELT** Project[36] is a free open platform for research and analysis of the global database. All the datasets released are free, open, and available for unlimited and unrestricted use for any academic, commercial, or governmental use. Also, it is possible to download the raw datafiles, visualize them, and analyse them at limitless scale. Recently, the

[33] https://sites.google.com/site/motazsite/corpora/osac

[34] https://www.sketchengine.co.uk/

[35] https://sites.google.com/view/arabicweb16

[36] https://www.gdeltproject.org/

GDELT Project has started to create linguistic resources. In fact, 9.5 billion words of worldwide Arabic news has been monitored over 14 months (February 2015 to June 2016) to make a trigram dataset for the Arabic language. Consequently, an Arabic trigram table of the 6,444,208 trigrams that appeared more than 75 times is produced[37].

It is worth mentioning that the drawbacks of the previous corpora have been addressed in the OSIAN corpus. I.e., unlike its trigram table, the corpus of GDELT is not available. The ArabicWeb16 corpus is free to download but it needs to be filtered, cleaned, and converted into an appropriate machine-readable form. The OSAC is free and cleaned, but it is very small compared to the other corpora, and created by downloading texts from websites unselectively with respect to their text type or content. Finally, none of these corpora is annotated.

3.2. Methodology and Tools

In this section, we describe the crawling and the annotation tasks as well as the tools used.

3.2.1. Data Acquisition

In a first step the data needs to be crawled from the World Wide Web. Since the crawled data are often duplicated, use different encoding, etc; they need to be cleaned and filtered. Therefore, the following processing steps were executed.

For crawling and processing, the typical procedures of the Leipzig Corpora Collection (LCC[38]) (Goldhahn et al. 2012; Quasthoff et al. 2014) were applied. The LCC started as "Projekt Deutscher Wortschatz[39]" in the Nineties as a resource provider for digital texts in the German language, mostly based on newspaper articles and royalty-free text material.

Today, the LCC offers corpus-based monolingual full form dictionaries in more than 200 languages mainly based on online accessible text material, classified by several criteria like the year of acquisition, text genre, country of origin and more. Since June 2006, in addition to direct access via a Web interface, LCC data is also offered for free download.

Normally, the **CURL-portal** (Crawling Under-Resourced Languages[40]) allows creating Web-accessible and downloadable corpora by simply entering URLs into the portal. However, for compiling the OSIAN corpus data, an adapted version of the **CURL-portal** of the LCC was utilized (Goldhahn et al. 2016). Expressly, the crawling was conducted in March 2018. The crawler was configured to only download web pages on the initial domains. Further, LCC applies a strict politeness policy, i.e., it respects entries in each server's "robots.txt", making certain websites' contents inaccessible. In addition, a delay of about 10 seconds before the same servers are queried again is applied. Consequently, six million URLs were crawled resulting in 148GB of compressed html-files in the Web ARChive (WARC) archive format.

The data of the OSIAN corpus have been drawn from 31 carefully selected and reliable Web domains around the world. The aim is to make the corpus balanced and to create country-specific sub-corpora, in addition to covering diverse topics and including high-quality texts.

[37] https://goo.gl/MZZkDJ

[38] http://corpora.uni-leipzig.de

[39] http://wortschatz.uni-leipzig.de

[40] http://curl.corpora.uni-leipzig.de/

Furthermore, the data processing was carried out according to the mostly language independent processing chain described in (Goldhahn et al. 2012): steps such as extracting raw text from the WARC file format, sentence separation, and cleaning of text using regular expressions are involved. Furthermore, since the crawler writes the data in one large file, we developed a tool for extracting the texts based on the Web domain. For each Web domain, the tool extracts and saves each article/page in a single file. Finally, these articles are assigned to the respective country. A list of the crawled Web domains, the number of articles extracted, and the countries covered are provided in the Appendix "C".

It is worth mentioning that the number of extracted articles varies from a Web domain to another. This may be explained by the fact that some domains were only restricted by the duration of the crawling and the low frequency of queries sent to the same server, which is the case where a large dataset was crawled, whereas, domains with few data ran out of crawlable URLs before the crawling finished. This could be due to "robots.txt" restrictions or there could have been links to other domains or similar, which we did not follow.

3.2.2. Corpus Annotation

As illustrated in previous chapters, lemma and PoS tags are among the widely used and important corpus annotation forms. Taken together, both these annotation forms are very beneficial and affect directly the performance of subsequent text analysis in NLP and IR. For both tasks we used the well-established Treetagger, which we adapted and evaluated previously for Arabic PoS tagging and lemmatization (see Chapter 4). It is worth mentioning that the adapted model was improved and retrained using new linguistic resources namely the Frequency Dictionary of Arabic (Buckwalter and Parkinson 2014). This frequency dictionary contains the top 5,000 words that were derived from a collection of representative corpora that include 30 million words of both written texts and transcribed speech.

3.3. Statistical Analyses

In this section, we highlight the characteristics of the OSIAN corpus using some statistical analyses. All in all, this corpus consists of about 3.5 million articles comprising more than 37 million sentences and roughly 1 billion tokens.

3.3.1. Word Length Statistics

The average length of words varies from 7 to 12 letters in many languages[41]. According to Mustafa (2012), the average length of Arabic words in a normal text is five letters. The following Table 5.3 displays the percentage of words covered in the OSIAN corpus with respect to their lengths for unique words and duplicated words.

Considering the whole corpus, 36% of the words have a length above six letters if duplicated words are included, the length of. For unique words the percentage is increased to 75%. This is possible due to the concatenation property using both affixes and clitics. Consequently, the OSIAN corpus offers a good ground to evaluate techniques that aim to reduce a word to its base such as stemming and lemmatization since 75% of its tokens is above six letters and the stemmed/lemmatized Arabic words normally consists of five letters as an average.

[41] http://www.ravi.io/language-word-lengths

Note that tokens with length superior to 10 letters are not considered since news articles contain phrases written without space characters between words as well as non-derived and concatenated words, such as "الأورومتوسطي"/Euro-Mediterranean, "الكهرومغناطيسية"/Electromagnetism, etc. This explains why we had more than 2 million unique tokens that consist of over 11 letters which is an irrational result for the Arabic language.

Table 5.3 Word length statistics

Word Length	Occurrence (Unique)	Percentage	Occurrence (Duplicated)	Percentage
2	4,180	0,03%	113,129,168	12,22%
3	45,723	0,28%	148,295,530	16,03%
4	412,528	2,52%	154,159,209	16,66%
5	1,550,485	9,48%	175,925,523	19,01%
6	2,877,426	17,59%	133,290,941	14,40%
7	3,353,777	20,50%	107,877,916	11,66%
8	2,864,584	17,51%	54,007,298	5,84%
9	1,919,115	11,73%	20,526,042	2,22%
10	1,196,370	7,31%	9,072,780	0,98%
>10	2,137,492	13,06%	9,050,623	0,98%
Total	16,361,680	100%	925,335,030	100%

3.3.2. Word Frequency List

Calculating word frequencies enables us to indicate the distribution of words across the text categories. Besides, it is feasible to produce word frequency lists using the tokens' PoS tags instead of their orthographic status. For those who face challenges in developing in-house tools to perform these and other analyses, there are some free and useful tools which can be relied upon such as LancsBox (Brezina et al. 2015) and Ghawwas (Almujaiwel and Al-Thubaity 2016).

Obviously, function words will be at the top of the frequency wordlist. Nevertheless, the words thematically organized in Table 5.4 are also among the most frequent words.

In the context of IR and corpus linguistics, many of the top frequently words have no value or effect on further analyses since they are typical in news articles; examples include "العالم" (World: F=1,182,181; R=37), "الحكومة" (Government: F=667,862; R=73), and "مفاوضات" (Negotiations: F=524,035; R=101). However, the words listed in Table 5.4 are a result of the circumstances of the Middle East in recent years, FIFA World Cup, and the Brexit, which make these words occur frequently in various world news. Using LancsBox to analyse the corpus data, it was possible to calculate frequencies of words that are obvious collocates such as "كأس العالم" (World Cup), "الاتحاد الأوربي" (European Union), and "البيت الأبيض" (White House). Moreover, it is

also possible to calculate statistical information about the association, the strength of collocation, and the comparative frequencies of word forms in the overall data of the OSIAN corpus or in country-separated data.

Table 5.4 Relevant words from the frequency wordlist

Theme	**Word**	**Frequency (F)**	**Rank (R)** (in whole corpus)
Persons	(Trump, President of USA) ترامب	608,176	81
	(Salman, King of Saudi) سلمان	380,086	164
	(El-Sisi, President of Egypt) السيسي	114,586	687
Countries	(Syria) سوريا	960,732	51
	(United Kingdom) بريطانيا	862,156	57
	(Qatar) قطر	704,457	70
Topics	(Election) الانتخابات	482,688	117
	(Brexit) بريكست	434,376	134
	(World Cup) كأس العالم	349,873	188
Organisations	(NATO) الناتو	387,174	161
	(European Union) الاتحاد الأوربي	177,383	448
	(White House) البيت الأبيض	124,762	648

3.4. Corpus Format

The XML-format is used to facilitate the use of the corpus. This is the first version of the OSIAN corpus which consists of separate directories for each country. Furthermore, each directory includes the articles in XML format, where the sentences are lemmatised and PoS tagged. Moreover, the XML files contain metadata to provide information about domain names, webpage location, and the date of extraction. For more illustration, Figure 5. 3 presents a sample of the XML files.

Note that some Web domains include in their URLs the topic of the published articles like the sample provided in Figure 5. 3 where the word “Science and tech” appeared in the article’s URL. This is another feature that can be used to classify the articles based on their topics, one among other techniques, to prepare them for classification and topic detection. Unfortunately, not all the URLs include such information; therefore, the topic label remains “unknown” till a solution is found (using topic detection and tracking methods).

```
<?xml version="1.0" encoding="UTF-8"?>
<Article num="1">
<Source name="BCC">
  <Date>2018-03-19</date>
  <Location>http://www.bbc.com/arabic/scienceandtech/2014/08/140829_smart_watches_sa
msung_lg</Location>
  <Topic> Science and Tech</Topic>
  <Language>ara</Language>
</Source>
<Text>
أعلنت شركتا سامسونغ وإلى جي الكوريتين الجنوبيتين طرح المزيد من الساعات الذكية…
</Text>
<Annotation>
  <Sentence id="1">
    <Word Surfaceform="أعلنت" PoS="VERB" Lemma="أَعْلَنَ" />
    <Word Surfaceform="شركتا" PoS="NOUN" Lemma="شَرِكَة" />
    <Word Surfaceform="سامسونغ" PoS="PN" Lemma="سَامْسُونْغ" />
    <Word Surfaceform="وإلى" PoS="PRT" Lemma="إلى" />
    <Word Surfaceform="جي" PoS="ABR" Lemma="جى" />
    <Word Surfaceform="الكوريتين" PoS="ADJ" Lemma="كُورِيّ" />
    <Word Surfaceform="الجنوبيتين" PoS="ADJ" Lemma="جَنُوبِيّ" />
    <Word Surfaceform="طرح" PoS="NOUN" Lemma="طَرْح" />
    <Word Surfaceform="المزيد" PoS="NOUN" Lemma="مَزِيد" />
    <Word Surfaceform="من" PoS="PRT" Lemma="مِنْ" />
    <Word Surfaceform="الساعات" PoS="NOUN" Lemma="سَاعَة" />
    <Word Surfaceform="الذكية" PoS="ADJ" Lemma="ذَكِيّ" />
    …
  </Sentence>
…
</Annotation>
</Article>
```

Figure 5. 3 A sample of OSIAN corpus encoded in XML format

3.5. CLARIN Integration

CLARIN[42] (**C**ommon **L**anguage **R**esources and **T**echnology **I**nfrastructure) is a European Research Infrastructure established in 2012 and took up the mission to create an online environment to provide access to language resources (in written, spoken, or multimodal form) for the support of scholars in the humanities and social sciences, and beyond (de Jong et al. 2018). Currently, CLARIN also offers advanced tools to discover, explore, exploit, annotate, analyse, and combine such data sets wherever they are located.

The CLARIN centre at the University of Leipzig, among others, is working on expanding available resources for a variety of languages with a dedicated focus on lesser-resourced ones. Unsurprisingly, a strong focus of CLARIN has been laid so far on resources for European languages. The integration of more data for non-European languages will broaden and extend possible research questions that users of the infrastructure can approach.

[42] https://www.clarin.eu/

Based on standard procedures and workflows that have been proven effective for "in-house" resources, the OSIAN corpus is processed, archived and published into the CLARIN infrastructure. This includes the publication of descriptive metadata via OAI-PMH[43], direct access to the plain text material, and integration into the WebLicht annotation platform and CLARIN's Federated Content Search FCS. In the future, the corpus will be made available via the KonText advanced corpus query interface for the Manatee-open corpus search engine (as used in the NoSketchEngine). This will enable compatibility with the FCS-QL specification v2.0 and will allow querying text and annotation layers such as part of speech and lemmas.

On the other hand, the current version and any updates of the OSIAN corpus can be found through our team website[44]. In addition, the corpus is available via the CLARIN research infrastructure, connecting them to central services such as VLO and FCS for metadata and content search. In the future, we will extend the OSIAN corpus to cover more international Arabic news with a diachronic and geographical perspective to make the corpus an ideal choice to explore language change and variation. Regarding CLARIN-integration, FCS 2.0 and the querying of annotation layers is planned to be supported.

4. MulTed Corpus

Given their importance, the demand for multilingual parallel resources is increasing primarily for those including under-resourced languages. However, the problem of building a balanced mix of multilingual texts in sufficient quantities and with a high-quality of translation becomes ever more central. This bottleneck becomes quite prohibitive when any further processing, such as sentence-alignment or PoS tagging, are to be involved. Therefore, building multilingual parallel corpora is becoming the focus of many NLP research fields. In this chapter, we introduce the MulTed corpus, a new multilingual aligned and tagged parallel corpus of subtitles extracted from TED talks. This corpus is designed for many NLP applications, such as statistical machine translation, language recognition, and bilingual dictionary generation, where the sentence-alignment, the PoS tagging, and the size of corpora are influential. Currently, the corpus has subtitles that cover 1,100 talks available in over 30 languages. Yet, the subtitles are classified based on a variety of topics such as Business, Education, and Sport. Regarding the PoS tagging, Treetagger is used and, to make the PoS tagging maximally useful, a mapping process to a universal common tagset is performed. Finally, we believe that the use of such a corpus can be a significant contribution to the literature of NLP and corpus linguistics, primarily for under-resourced languages.

4.1. The Value of the MulTed Corpus

Regardless of the difficulty of building multilingual corpora, they are very valuable for many applications in NLP field (Tiedemann 2007). Statistical Machine Translation (SMT) is one such application that has achieved significant impact with the help of such corpora. The main challenge for the success of the translation process is the access to high-quality training data. SMT can ease the task for translators, by providing an initial translation, which can be later post-edited (Green et al. 2013). Basically, SMT systems are generally trained using sentence-aligned parallel corpora (Brown et al. 1993; Callison-Burch et al. 2004). The use of a sentence-aligned

[43] See for example http://hdl.handle.net/11022/0000-0007-C65C-3

[44] http://oujda-nlp-team.net/en/corpora/osian-corpus/

corpus for machine translation is a pattern matching process because each component of the source sentence is matched to the targete one (Le et al. 2012). To retrieve translation results, this method does not require syntactic and semantic analysis. However, the accuracy depends on the number of sentences stored in the parallel corpus. Therefore, a new multilingual corpus like the MulTed corpus could contribute to reducing the cost of creating training data for new language pairs and domains in order to increase the performance of their SMT systems. Recently, the overlapping of corpus linguistics and descriptive translation studies has contributed to the birth and rise of corpus-based translation studies (CTS). CTS has become a major paradigm and research method. It applies statistical analysis of words or phrases in parallel corpora in different languages to obtain probabilities of translations. Furthermore, *Hu* (2016) has explained in more detail how parallel corpora can be used in translation teaching; primarily on the establishment of a corpus-based mode of translation teaching and the use of corpora in compiling translation textbooks.

PoS tagging and its natural successor, parsing, are vital tasks in NLP and corpus linguistics. Knowing that different ambiguity patterns are likely to occur in different places across languages, then, combining information from many languages creates a clearer picture of each language (Naseem et al. 2014). When a parallel corpus is available, cross-lingual PoS tagging can be used to assess the effectiveness of cross-linguistic projection of morphological features to an under-specified target language (Sylak-Glassman et al. 2015). Similarly, with the help of parallel treebanks, syntactic annotation may achieve a notable impact (Xing et al. 2016), as for example, the syntactic annotation applied using the parallel corpus "Prague Czech-English Dependency Treebank" (Bojar et al. 2012; Hajic et al. 2012). A parallel corpus will be more valuable for some NLP tasks if it is aligned at the level of words as well as sentences. For example, in Word Sense Disambiguation process, the word senses can be derived from word alignments on a parallel corpus instead of a pre-defined monolingual sense-inventory such as WordNet (Lefever et al. 2011). Furthermore, exploiting the text and structure of a parallel corpus (e.g., Wikipedia) provides enormous multilingual training annotations for Named Entity Recognition (Bodnari et al. 2013; Nothman et al. 2013).

A combination of multilingual corpora and the query translation can also be used to enhance the performance of the Cross-Lingual Information Retrieval (CLIR) (Bhattacharya et al. 2016). CLIR is a task used to search and retrieve the relevant information required, where "source" documents are written in a language while the user's queries are in another one. CLIR models can be trained with document-aligned parallel corpora, or they can include a translation mechanism followed by monolingual Information Retrieval. This method can lead to better retrieval effectiveness (Magdy and Jones 2014).

4.2. State of the Art

In this section, we present some relevant bilingual and multilingual parallel corpora. Since their construction is expensive in terms of time and effort, only few corpora are freely available.

4.2.1. Bilingual Parallel Corpora

In last decade, a range of bilingual parallel corpora has been established especially those including the English language. For instance:

1) The **CzEng** corpus (Dušek et al. 2012) is a Czech-English parallel corpus and freely available for non-commercial research or educational purposes. Several features have

been included in the last release such as morphological tags, surface syntactic, and automatic co-reference links. Most sources of this corpus are books, EU Legislation, Movie Subtitles. By increasing its size, the corpus reached 15 million sentence pairs (about 200 million tokens per language).

2) The **Scielo** corpus (Neves et al. 2016) is a freely available parallel corpus of scientific publications for the biomedical domain (biological sciences and health sciences). The corpus data have been retrieved from the Scielo database. The corpus is available for three language pairs: Portuguese-English (about 86,000 documents in total), Spanish-English (about 95,000 documents) and French-English (about 2,000 documents).

3) A parallel corpus, like Scielo, has been constructed in two language pairs, Portuguese-English and Portuguese-Spanish, based on scientific news texts. These texts have been crawled automatically from the multilingual Brazilian magazine Pesquisa FAPESP Online; then, aligned at the document and sentence level. The corpus contains about 2,700 parallel documents totalling over 150,000 aligned sentences per language (Aziz and Specia 2011).

4) Other bilingual parallel corpora have been developed, such as the Persian-English corpus (Mohammadi and GhasemAghaee 2010), the French-English corpus (Germann 2001), and the Japanese-English corpus (Utiyama and Isahara 2007). Most sources of these bilingual corpora covered restricted topics such as legislation (e.g., debates of the European parliament), administration documents, and technical documentation like operating system software manuals.

4.2.2. Multilingual Parallel Corpora

Concerning multilingual parallel corpora, **OPUS** is probably the largest collection of freely available parallel corpora in different languages with a considerable size and variety (Tiedemann 2012). For example, it contains the **EuroParl** (Koehn 2005) and the **JRC-Acquis** (Steinberger et al. 2006) corpora. These two corpora both contain the European Union (EU) documents of mostly legal nature such as the proceedings of the European Parliament. Both corpora are also available with bilingual alignments in all language pairs, including English. However, the EuroParl exists only in eleven European languages and contains none of the languages of the new Member States or of a candidate country. Altogether, this corpus contains about 30 million words for each of the 11 languages. On the other hand, the JRC-Acquis is available in 21 official EU languages with an average size of roughly 9 million words per language. A first version of the United Nations Parallel Corpus (Ziemski et al. 2016), that is similar to the **MultiUN** corpus, is composed of official records and other parliamentary documents of the United Nations that are in the public domain. This corpus contains documents that were produced between 1990 and 2014 and manually translated into the six official languages of the United Nations.

Most mentioned corpora are restricted only to the high-density languages such as English and the European languages. Unfortunately, some languages are not included in a considerable number of relevant multilingual parallel corpora. For instance, the Arabic language is covered by a small number of bilingual and multilingual corpora such as the tiny Arabic-English parallel corpus (10K sentences) used to build an Arabic stemmer based on statistical machine translation using an English stemmer (Rogati et al. 2003). Another Arabic-English parallel corpus has been adopted to handle the word translation disambiguation (Ahmed and Nürnberger 2008). In addition, a multilingual named entity corpus for Arabic, English, and French has been developed

based on comparable newswires from the "Agence France Presse" covering the period 2004-2006 (Mostefa et al. 2009). Finally, a free Arabic-English parallel corpus has been built within the project **MEDAR** (Maegaard et al. 2009) (MEDiterranean ARabic language and speech technology), supported by the European Commission's ICT program and which has been running from 2008 to 2010. The project addressed international cooperation with the Arabic region on Human Language Technology.

The following works are much more related to the MulTed:

1) **SwissAdmin** (Scherrer et al. 2014): It is one of the fewer freely available multilingual and PoS tagged parallel corpora. It is built of press releases from the Swiss Federal Administration. The corpus is available in four languages (English, German, Italian and French). It is released in three versions: plain texts of approximately six to eight million words per language, sentence-aligned bilingual texts for each language pair, and a PoS tagged version. The annotation has been performed automatically by the ***Fips*** multilingual parser (Wehrli and Nerima 2015).
2) **AMARA** corpus (Abdelali et al. 2014): It is a parallel corpus of educational video subtitles, multilingually aligned for 20 languages, i.e., 20 monolingual corpora and 190 parallel corpora. The data of this corpus were collected in cooperation with the Amara platform[45] using an in-house crawler. 3,000 videos have subtitles available in at least six languages and 1,000 videos have subtitles available in 25 languages. However, the corpus is not PoS tagged and is not freely available.
3) **"WIT^3"** project (Cettolo et al. 2012), an acronym for **W**eb **I**nventory of **T**ranscribed and **T**ranslated **T**alks, is a collection of lecture translations that have been automatically crawled from the TED talks in a variety of languages. The purpose of this project is to support the machine translation evaluations campaigns of the International Workshop on Spoken Language Translation (IWSLT) (Paul et al. 2010). As of October 2011, 17 thousand transcripts corresponding to translations of around 1,000 talks have been collected. The corpus of the "WIT^3" project is likely to be a haphazard collection of subtitles that are not balanced or classified based on the variety of TED topics. It is also not PoS tagged.

Based on what has been described above, we propose a new corpus to address the limits of the mentioned corpora. I.e., a freely available multilingual parallel corpus, sentence-aligned, PoS tagged, covering under-resourced languages from different families, and well balanced in terms of domains and topics.

4.3. Data Collection Procedure

In this section, we present the data resource and the tools used for the collecting and filtering processes. Typically, Internet users provide subtitles in various languages voluntarily. Thus, huge online databases for subtitles are available for free on the web. Sometimes, translators provide different subtitles versions of the same language for the same videos. What's more, subtitles are different from other parallel resources in various aspects, since most of them are transcriptions of spontaneous speech. Thus, they can easily be linked to the actual sound signals (Tiedemann 2007). One of the most relevant and available subtitles on the web are those provided for TED talks.

[45] http://amara.org

4.3.1. TED Talks

"TED talks" is a library of talks, filmed at independently non-profit organized events in over 130 countries. Due to the popularity of TED events worldwide, presenting high-quality content on many different topics, amazing efforts have been undertaken by at least 25,000 volunteers to generate about 40,000 translations into 101 languages or more (Abdelali et al. 2014). The TED website[46] makes the video recording of the best talks and all their subtitles available under the Creative Commons BYNC-ND license. The talks are presented in an excellent and original style by very skilled speakers who cover a wide variety of topics under the slogan of "Ideas worth spreading". The talks are divided according to the languages, topics, countries and posted dates. As for the translation process, using TED talks imply dealing with spoken language, which is structurally less complex, formal and fluent, than written language. Further, the translators are not required to translate literally, but they have to follow the structure and the rhythm (i.e., timing) of the English, as it is explained in the TED platform[47], to avoid the usual rephrasing and reordering tasks in the ordinary translation of written documents. For instance, a subtitle must not contain the end of one sentence and the beginning of another, it should be synchronized with the talk, unless the duration of a subtitle must be extended for a good reading speed.

One of the reasons that we use TED subtitles, is the high-quality of their translation. As reported on its platform[48], the subtitles go through the following steps before publication:

1. **Transcription:** The TED platform provides an original transcript;
2. **Translation:** Subtitles are translated from the original language into the target language, using a simple online interface;
3. **Review:** Subtitles are reviewed by an experienced volunteer (someone who has subtitled 90 minutes of talk content);
4. **Approval:** Before publication, reviewed translations are approved by a TED Language Coordinator or staff member.

4.3.2. Data Collection Tools

For many languages, the small number of volunteers cannot keep up with the fast pace in which new content is appearing on the TED website. Thus, we did not collect all the list of available videos. The crawling yielded over 30,000 translations, corresponding to 1,100 videos in 101 different languages. The initial collection was completed between May 10 and June 20th, 2016.

For the data collection, Google2SRT[49] is used to retrieve subtitles automatically in SRT file (.srt) format. Google2SRT is a freely available tool which allows downloading, saving and converting multiple subtitles and translations from YouTube and Google Video to SubRip format. Google2SRT can extract subtitles from XML files as well as from a direct video's hyperlink or a list of video URLs saved in a TXT file. One of its useful features is the ability to

[46] http://tedxtalks.ted.com

[47] http://translations.ted.org/wiki

[48] https://www.ted.com/participate/translate/get-started

[49] https://sourceforge.net/projects/google2srt/

select the translations that include multiple versions of the same language. It also allows choosing and saving the preferred languages in one folder in addition to the original transcript. The crawler *HLTWebManager* could also be used for subtitles extraction (Cettolo et al. 2012), however, its use leads to an additional process to collect and link the original transcript to its translations.

4.3.3. Filtering and Topic Classification

TED talks cover a wide range of domains and topics, but not all videos come with a considerable number of translations. Thus, only the resources that met our criteria have been selected. To do so, we manually:

- Selected talks with subtitles of more than specific 15 languages, especially poor or medium density languages. We also made sure to select languages from different families such as the six official languages of the United Nations which are Arabic, Chinese, English, French, Russian and Spanish.
- Selected and organized the talks from a variety of topics to ensure heterogeneity and equilibrium in the corpus. We choose the 11 following topics: “Architecture and Design”, “Art and Creativity”, “Culture and Stories”, “Economic and Innovation”, “Education and Learning”, “Global Issues”, “Health and Medicine”, “Nature and Environment”, “Science and Tech”, “Social Issues” and “Sports and Adventure”. From each topic, 100 videos were selected.

4.4. Sentence-alignment Methods

The aim of this section is to present methods that attempt to handle sentence-alignment. To build aligned parallel resources, several methods are proposed, and tools have been developed for this serious process. For instance, *Moore* (2002) has presented and discussed relevant alignment methods namely sentence-length-based and word-correspondence-based methods. As stated by *Moore*: “the sentence-length-based methods are relatively fast and fairly accurate”; whereas, “word-correspondence-based methods are generally more accurate but much slower, and usually depend on cognates or a bilingual lexicon”. Consequently, he has proposed a new method that combines these two methods in order to achieve a high accuracy level at a low computational cost and without requiring any knowledge of the languages or the corpus beyond division into words and sentences. Similarly, *Yang* and *Li* (2004) have investigated and compared some length-based and text-based methods. Furthermore, *Vandeghinste* and *Sang* (2004) have applied an alignment method based on lexicalized similarities to align a parallel transcript corpus for sentence compression. Based on Wikipedia, another method has been adapted that uses an extended link-based bilingual lexicon to build a Farsi-English parallel sentence-aligned corpus (Mohammadi and GhasemAghaee 2010). Finally, a sentence-alignment method based on maximum entropy model using anchor sentences has been proposed since other methods might not be accurate for peculiar languages, such as Chinese (Che et al. 2016).

In additon, some scientific groups work on word alignment level using standard tools, especially for statistical machine translation systems; such as **GIZA++** (Tian et al. 2011), **fast align** (Dyer et al. 2013), and **efmaral** (Östling and Tiedemann 2016).

Despite the existence of a number of alignment methods and tools, finding a suitable to aligning the MulTed corpus is challenging due to the diversity of its languages. A sentence-

aligner tool like **YASA** (Lamraoui and Langlais 2013) could perhaps be used in this work. YASA is an open source sentence aligner tool that seems to be relatively language independent, or easy to adapt to a variety of language pairs. However, the MulTed data are aligned at both the segment and sentence levels, which are more suitable for subtitles at least at this stage.

4.5. Sentence-alignment Procedure

Typically, when the data are harvested, they are probably noisy. I.e., subtitles could contain wrong or incomplete components. In fact, an investigation confirms that SMT systems are highly tolerant to noise, and the performance degrades seriously only at very high noise levels (Goutte et al. 2012). Consequently, cleaning the noisy parallel data by detecting and removing incorrect alignments can improve the performances. As a result, the content of collected subtitles has been reviewed carefully using the attributes in the SRT file which include the talk and sentence IDs as well as the time-slot which is the start and end times of the segment.

Since all subtitles are segmented based on sound, there is one "segment ID" for every caption that appears on the screen at a specific timeframe. Consequently, the first method is based on time slots segmentation, and the content of subtitles is segment-aligned using the segment "IDs". Next, many heuristic checks are performed to assess the alignment. I.e., a subtitle is discarded if either the sequences of segments "IDs" or the total number of segments differ from those of the English transcriptions, resulting in about 2% of the subtitles being eliminated. A sample of the remaining subtitles (e.g., English-Arabic and English-French) has been manually checked to be sure that are successfully segment-aligned.

Note that a segment could be either a whole sentence or a part of it. Considering that the subtitles contain proper punctuations, a second alignment method is applied using the punctuation marks as boundaries to form a complete sentence. In doing so, the sentences are regenerated by concatenating on both sides consecutive segments until a strong punctuation mark (e.g., period and question mark) is detected on the target side. Finally, the entire corpus is sentence-aligned considering English as a pivot language and the average number of sentences obtained is about 1.5 million per language. Figure 5.4 presents a sample for more illustration.

4.6. PoS Tagging

Unlike the sentence-alignment procedure that includes all the 101 languages covered in the MulTed corpus, the PoS tagging process involved only 30 languages. Treetagger is probably the most widely used language independent PoS tagger, and has successfully annotated texts in more than 30 different languages with an average accuracy of 95% (e.g., English 96.36% (Schmid 2013), German 97.53% (Schmid 1999), Russian 97.31% (Kotelnikov et al. 2017), Classical Latin 95.5% (Field 2016), and Arabic 94.7% (Imad Zeroual, Lakhouaja, et al. 2017)). A sample of 500,000 words of the Arabic part of the MulTed corpus has been used to evaluate the performance of Treetagger (Zeroual and Lakhouaja 2017). The reported accuracy rate is 88.87% (see Chapter 4, Section 5.6).

Since the Treetagger is separately adapted to different languages, the used tagsets for each language are not identical. To annotate the MulTed corpus with a common set of tags, the universal tagset (see Chapter 4, Section 4.1) was adopted. As observed, all these 30 languages share 12 sets of basic tags. Since EAGLES recommendations for the morphosyntactic annotation

of corpora report that there are 13 main categories[50] considered obligatory for most languages, the mapping of the tagsets used for each language was done manually to convert the subcategories to the main categories.

```
<?xml version="1.0" encoding="UTF-8"?>
 <Talk id="Fg_JcKSHUtQ">
  <Category>Architecture and Design</Category >
  <Title>A robot that flies like a bird</Title>
  <Speaker>Markus Fischer</Speaker>
  <Time-slot>00:00:15,259 --> 00:00:18,259</Time-slot>
    <Segment id="1">
      <Original_text Lang_id="en">It is a dream of mankind, </Original_text>
      <Translation Lang_id="fr" Translator="Elisabeth Buffard" Reviewer="Alban Lefebvre">
      C'est un rêve de l'humanité, </Translation>
      <Translation Lang_id="es" Translator="Veronica Martinez Starnes" Reviewer="Sebastian
      Betti">Es un sueño de la humanidad, </Translation>
      <Translation Lang_id="ar" Translator="Faisal Jeber" Reviewer="Anwar Dafa-Alla">
      هو حلم البشرية،</Translation>
      <Translation Lang_id="zh-TW" Translator="Chunxiang Qian" Reviewer="Angelia King">
是人类的一个梦想。</Translation>
      <Translation Lang_id="th" Translator="Heartfelt Grace" Reviewer="Paravee Asava-
      Anan"> เป็นความใฝ่ฝันของมนุษย์, </Translation>
      …
    </Segment>
    …
 </Talk>
```

Figure 5.4 A sample of a segment-aligned subtitle encoded in XML format

4.7. Statistical Information

The current version of the MulTed corpus has subtitles covering approximately 1,100 talks available in over 30 languages. The corpus comprises 30,000+ subtitles that contain 7.6 million aligned sentences with altogether over 46 million tokens. We sentence-aligned the entire corpus considering English as a pivot language, i.e., the alignment is done between English and the other languages. Then, the subtitles were classified manually into 11 categories based on the variety of TED topics. Finally, after collecting, filtering, sentence-aligning, and tagging the data, we were left with the following database presented in Table 5.5.

Table 5.5 General information about the MulTed corpus

	No. of talks	Number of languages	Number of subtitles	Number of segments	Number of tokens
Total	1,100	101	30,057	7.6 million	46+ million

To estimate the number of segments and tokens, the following tasks were performed:

[50] http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#cmobli

- **Normalization:** to remove non-needed subtitle components (i.e. time slots and talk id...) and special characters except punctuation marks;
- **Tokenization:** to break up the text into individual tokens using as delimiters, white-space, and newline.

In terms of languages, the top 30 languages have subtitles in a considerable number of videos (i.e., at least 500). What's more, many resource-poor languages are covered in the corpus such as Thai (**th**) with 624 subtitles and Indonesian (**id**) with 509 subtitles. Figure 5.5 Figure 5.5 Distribution of the top 30 languages by number of talkspresents the overall distribution of these top 30 languages by a number of talks.

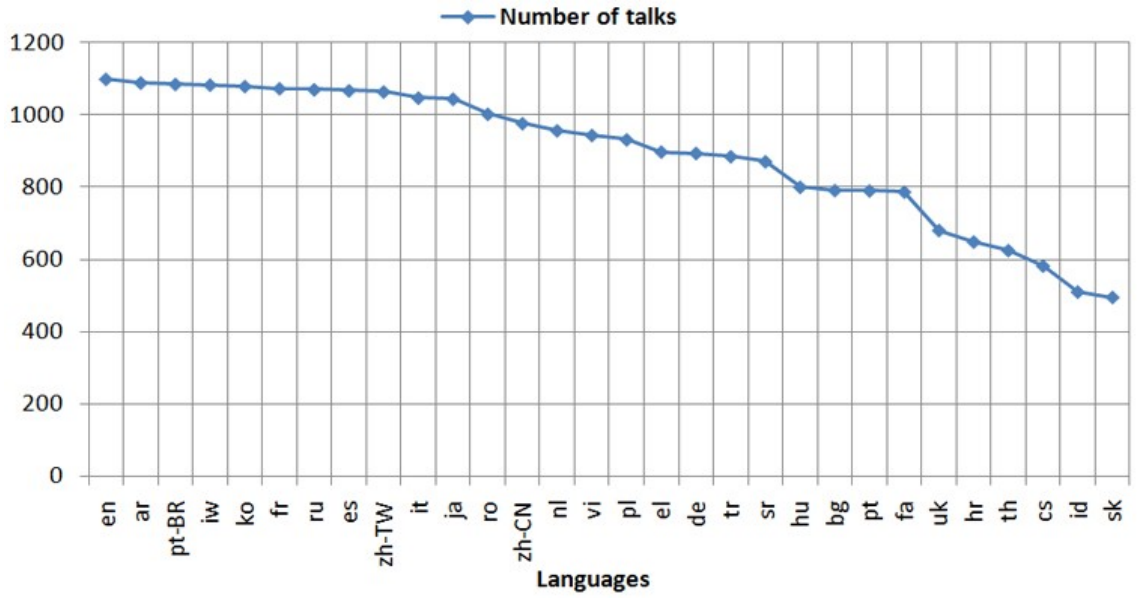

Figure 5.5 Distribution of the top 30 languages by number of talks

Next, we present more details about the top 15 languages (the other 15 languages are included in Appendix "D"). What's more, Table 5.6 displays the number of monolingual files, the number of tokens per language, and the number of segments pairs with English, respectively.

4.8. Corpus Format

In addition to the original format, XML, as an encoding language, is used to facilitate the use of the corpus. Thus, this corpus is released in three versions:

1. Plain text version of approximately 0.6 (e.g., Slovenian) to 2.1 (e.g., English) million tokens per language;
2. Sentence-aligned bilingual texts version for each language pair;
3. PoS tagged version for each talk.

Moreover, the XML files in all versions contain tags and attributes to provide further metadata of the talk. For instance, talk id, title, category, translator, time slot, speaker, and language id. Their meaning is self-explanatory. Figure 5.6 exhibits a sample of the PoS tagged version of an Arabic subtitle (another sample of English is given in Appendix "D").

Table 5.6 The number of segments pairs with English

Languages	Nb. of files	Nb. of tokens	Nb. of segments
English	1,100	2,134,155	275,847
Arabic	1,090	1,703,114	271,246
Portuguese-BR	1,085	1,978,894	270,255
Hebrew	1,085	1,568,768	270,534
Korean	1,080	1,417,473	266,449
French	1,074	2,124,021	267,901
Russian	1,071	1,684,719	270,825
Spanish	1,068	1,968,660	266,847
Chinese (TW)	1,065	326,739	265,535
Italian	1,048	1,859,765	261,477
Japanese	1,045	406,964	264,610
Romanian	1,003	1,750,662	242,909
Chinese (CN)	977	298,569	258,122
Dutch	957	1,726,997	240,410
Vietnamese	944	2,279,456	232,153

```
<?xml version="1.0" encoding="UTF-8"?>
 <Talk id="Fg_JcKSHUtQ">
  <Category>Architecture and Design</Category >
  <Title>A robot that flies like a bird</Title>
  <Speaker>Markus Fischer</Speaker>
  <Translation Lang_id="ar" Translator="Faisal Jeber" Reviewer="Anwar Dafa-Alla">
  <Time-slot>00:00:15,259 --> 00:00:18,259</Time-slot>
    <Segment id="1">
      <Word PoS="PRON" Lemma="هو">هو</Word>
      <Word PoS="NOUN" Lemma="حلم">حلم</Word>
      <Word PoS="NOUN" Lemma="بشريّ">البشرية</Word>
      <Word PoS="SENT" Lemma="unknown">,</Word>
    </Segment>
    …
  </Translation>
 </Talk>
```

Figure 5.6 A sample of a PoS tagged version of an Arabic subtitle

A related point to consider is that the parallel corpora cited in this chapter are still useful, and readers can benefit from them in a way or another. The purpose of building the MulTed

corpus is not to replace any of them, but to address some issues that could be for the benefit of researchers. In this section, we highlight the drawbacks of some existing parallel corpora, including the MulTed corpus as well as the advantages of this latter.

4.9. Discussion

The major drawback of many multilingual parallel corpora is probably that they are compiled based mainly on legislative or technical raw data (e.g., SwissAdmin and MultiUN). Also, these corpora are restricted mostly to high-density languages such as English and European languages (e.g., EuroParl). Besides, some of them are in fact bilingual pairs rather than multilingual (e.g., CzEng). Regarding those corpora that consist of subtitles, most of them target movies and TV shows subtitles. This creates a challenge for sentence-alignment due to the possibility of multi-speakers in the same time slot. Further, the language used in movies is informal. Thus, movies subtitles are not well-suited for machine translation purpose (Gelbukh 2011), and translating this language is much more challenging, since research in SMT is mostly based on the formal translation tasks (van der Wees et al. 2016). In addition, several parallel corpora are neither aligned nor PoS tagged. The MulTed corpus is not aligned on a word level, and the remaining 71 languages, such as Greek and Indonesian, are not PoS tagged.

Tiedemann et al. (2016) confirms that training data is the most effective way to increase translation performance. Therefore, the MulTed corpus has some significant advantages such as the high-quality of its translations because the crowdsourced transcriptions and the translations are reviewed by experienced translators, then, approved by TED Language Coordinators or staff members. Additionally, the TED talks are presented in well-structured language which makes this kind of corpus very valuable to build SMT systems. Indeed, since 2011, the transcriptions and the translations of TED talks are used yearly, as training and testing data, for an open evaluation campaign on spoken language translation in the International Workshop on Spoken Language Translation (IWSLT[51]). Furthermore, the MulTed corpus covers a variety of topics in the used raw database. Unlike "WIT^3", the subtitles of the MulTed corpus are classified and balanced manually into 11 categories based on these topics. Moreover, this corpus is characterized by language diversity since it covers high, medium, and poor density languages from different families. Besides, the MulTed is based on talks that have only one speaker, which helps the alignment process as well as text compression and summarization studies, as done in the European projects MUSA and Flemish ATraNoS to summarize the discussion of TV shows (Daelemans et al. 2004). Finally, it is a multilingual parallel corpus that covers over 30 languages, sentence-aligned, and PoS tagged.

5. Conclusion

This chapter has shed light on the procedure of building three different Arabic corpora: the Al-Mus'haf corpus, the OSIAN corpus, and the MulTed Corpus. The characteristics and the features of each corpus are presented and evaluated with the stat-of-the-art corpora. Further, all data have been stored in suitable format, such as XML, and will be publicly available to use for various research purposes. Expressly, the Al-Mus'haf corpus covers the Quranic Arabic text, where all the words are annotated with rich and important morphosyntactical information namely stem, stem pattern, lemma, lemma pattern, and root. These notable features have distinguished

[51] http://iwslt.org

this corpus from similar corpora. The MulTed corpus is a multilingual and PoS tagged parallel corpus with bilingual sentence-alignment and English as a pivot language. The corpus was made derived from the TED talks, where volunteers contribute transcriptions and translations that are available to the public. The corpus currently contains subtitles that cover 1,100 talks, including a variety of domains and topics. Besides, it is characterized by language diversity where at least 30 languages are well covered. Finally, a web-derived corpus called the OSIAN corpus was compiled based on 31 different international Arabic news broadcasting platforms. With a server-friendly crawling policy, we extracted one million web pages. In the future, we are particularly interested in extending the size of some corpora with a diachronic and geographical perspective to make the corpora suitable to explore a language change and variation. Additionally, we aim to develop new types of corpora such as learner and historical corpora.

The corpora developed have followed, as much as possible, the criteria mentioned in the first chapter. For instance:

- ✓ the corpora are well-defined and clearly described:
- ✓ they include freely available sources and all copyrights are respected;
- ✓ they are compiled in suitable machine-readable forms with a considerable size;
- ✓ they have been annotated using appropriate tools and forms taking into consideration the standardization aspects;
- ✓ the balance and representativeness are, to some extent, addressed while selecting data sources and topics.

It is necessary to emphasize that these three corpora developed do not fully represent the Arabic language. However, the prime motivation for building the corpora presented in this chapter is the lack of free Arabic corpora that can cope with the perspectives of Arabic NLP and IR, among other research areas. Furthermore, we expect that those corpora can be used to answer relevant research questions in corpus-based studies. Finally, one of our main objectives will be always finding new ways to improve the accuracy of the processing and the annotation tools as well as to adopt new and meaningful forms of annotation.

CHAPTER 6: Conclusion and Future Directions

1. Introduction

Corpora have intrigued linguists for centuries. From a historical view, previous works have shown that the Arabic lexicographers were among the first linguists that assembled large collections of Classical Arabic texts to create the earliest known dictionaries of any language. Since then, much more collections of texts have been compiled and used for different purposes. Starting from the 1960s onwards, the name "corpus" has been used to represent these collections. A few years later, fundamental criteria and instructions have been proposed by relevant corpora builders to follow in the design and the compilation of a general corpus.

Both Classical and MSA are understudied in computational linguistics and NLP, relative to its worldwide reach as the language of the Quran, its proud heritage, and lexical richness. However, for Arabic as for many other languages, most researchers have often built corpora and develop tools that may only suit their personal objectives and for a specific time without considering the design criteria and the language features to create well-defined corpora and robust tools that meet the standards.

During this thesis, our concern was initially to build the most needed types of Arabic corpora considering the Arabic language features but also in compliance with corpus linguistics standards. For this reason, the decision was made to investigate different methods used in previous works, develop and propose new tools and resources, adapt relevant tools to deal with the Arabic language texts, and conduct several experiences and comparative studies to provide a comprehensive description of the main stages followed while building corpora.

2. Summary of Contributions

This thesis presents novel contributions to the literature through the following aspects:

- ✓ Literature: a historical timeline of corpora building that goes back till the 8th century was provided. Besides, we conducted a survey that covers 100 well-known and influential corpora and presents a summarisation of data sources and different compilation methods used in relation to corpus characteristics like size and time consumed during the compilation process.
- ✓ Stemming: related concepts and terms to this task, that differ from its corresponding used in western languages, were clarified with numerous examples. Then, we developed a new Arabic stemmer and it was evaluated and compared to the state-of-the-art tools.
- ✓ Lemmatization: we studied and investigated the state-of-the-art of available tools (Al Khalil lemmatizer and Madamira), then, comparative and usability tests are performed.
- ✓ PoS tagging: we proposed a standard tagset considering the Arabic language features. Further, we carefully collected linguistics resources to create the required dictionaries to adapt and enhance the performance of three language-independent PoS taggers (Treetagger, TnT, and SVMTool).
- ✓ Corpora: we built three corpora (Al-Mus'haf, OSIAN, and MulTed) that cover both Classic and Modern Standard Arabic. Detailed information about the building

procedures and the characteristics of the constructed corpora are presented. Furthermore, they are compared to similar corpora, stressing their significant contribution to the literature.

Finally, the thesis is based on sixteen publications that make significant contributions to knowledge relating to building, processing, and annotating Arabic corpora. The published papers, the tools, and the corpora developed during this time are available on the team website (http://oujda-nlp-team.net/).

3. Publications

Most of the chapters of this thesis are based on the following publications:

Chapter 2:

1. Zeroual, I., & Lakhouaja, A., "Data science in light of natural language processing: An overview." Procedia Computer Science 127 (2018): 82-91. **DOI**: 10.1016/j.procs.2018.01.101.
2. Zeroual I., Lakhouaja A. (2018) Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go. In: Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham. **DOI**: 10.1007/978-3-319-67056-0_29
3. Zeroual, I., & Lakhouaja, A., "Stages in corpus building: A survey", The 6th International Conference on Arabic Language Processing, October 11-12, 2017, Fez, Morocco

Chapter 3:

1. Zeroual, I., & Lakhouaja, A., "Arabic Information Retrieval: Stemming or Lemmatization?" The 2nd International Conference on Intelligent Systems and Computer Vision (ISCV'17), April 17-19, 2017, Fez, Morocco. **DOI**: 10.1109/ISACV.2017.8054932
2. Zeroual, I., Boudchiche, M., Mazroui, A., Lakhouaja, A., "Developing and performance evaluation of a new Arabic heavy/light stemmer", The 2nd International Conference on Big Data, Cloud and Applications (BDCA'17), March 29-30, 2017, Tetuan, Morocco. **DOI**: 10.1145/3090354.3090371
3. Zeroual, I., Boudchiche, M., Mazroui, A., Lakhouaja, A., "Improving Arabic light stemming algorithm using linguistic resources", 2nd National Doctoral Symposium in the field of Arabic Language Engineering, (JDILA'15), October 28-29, 2015, Fez, Morocco.
4. Zeroual, I., & Lakhouaja, A., "Clitic Stemmer: A new Stemmer for Arabic language", 1st National Doctoral Symposium in the field of Arabic Language Engineering, (JDILA'14), February 8, 2014, Rabat, Morocco

Chapter 4:

1. Zeroual, I, and Lakhouaja A., "A Comparative Study of Standard Part-of-Speech Taggers." In: Ezziyyani, M. (Ed.), Advanced Intelligent Systems for Sustainable Development (AI2SD'2018): Volume 5: Advanced Intelligent Systems for Computing Sciences, Advances in Intelligent Systems and Computing, vol 915. Springer International Publishing. **DOI:** 10.1007/978-3-030-11928-7_75
2. Zeroual, I, and Lakhouaja A., "Feature-rich PoS Tagging through Taggers Combination: Experience in Arabic." Transactions on Machine Learning and Artificial Intelligence 5.4 (2017). **DOI**: 10.14738/tmlai.54.2981
3. Zeroual, I., Lakhouaja, A., and Belahbib R., "Towards a standard part of speech tagset for the Arabic language". Journal of King Saud University – Computer and Information Sciences, 2017, **DOI**: 10.1016/j.jksuci.2017.01.006
4. Zeroual, I., and Lakhouaja, A., "Adapting a decision Tree based Tagger for Arabic", 2nd International Conference on Information Technology for Organizations Development, March 30 – April 1st, 2016, Fez, Morocco. **DOI**: 10.1109/IT4OD.2016.7479306

Chapter 5:

1. Zeroual, I., and Lakhouaja, A., "MulTed: A multilingual aligned and tagged parallel corpus", Applied Computing and Informatics (2018). **DOI**: 10.1016/j.aci.2018.12.003
2. Zeroual, I., and Lakhouaja, A., "A new Quranic Corpus rich in morphosyntactical information", International Journal of Speech Technology, 2016, **DOI**: 10.1007/s10772-016-9335-7
3. Zeroual, I., & Lakhouaja, A., "Towards a Multilingual Aligned Parallel Corpus", 1st International Conference of High Innovation in Computer Science (ICHICS'16), June 01-03, 2016, Kenitra, Morocco.
4. Zeroual, I., & Lakhouaja, A., "Al-Mus'haf Corpus: A New Quranic Corpus rich in Morphosyntactical Information and accurate Part of Speech tagging", International Workshop on Computers and Information Sciences, (WCIS 2015), October 7-8, 2015, Tabuk, Saudi.
5. Zeroual, I., & Lakhouaja, A., "A New Quranic Corpus rich in Morphological Information", 5th International Conference on Arabic Language Processing, (CITALA 2014), November 26-27, 2014, Oujda, Morocco.

4. Limitations

Although the need is great to build Arabic corpora, especially reference and gold standard corpora, several factors can explain the limits of this study as well as most Arabic corpora building projects. It is well known that the creation of valuable corpora is expensive, time-consuming, and requires specialized personnel. Therefore, among the main challenges is the absence of funding and investment for the development of large and manually annotated Arabic corpora. However, launching such huge projects needs the collaboration of different national

institutions primarily Universities and research centres as well as the governmental funding and support.

In contrast to the rich-resourced languages, there is a lack of available tools and robust analysers that can deal effectively with the richness of morphology and syntax for both CA and MSA forms. Consequently, most Arabic corpora builders have often developed tools and annotation forms that comply with their own objectives without considering standardization and international aspects. During this thesis, we attempted to address this challenge and propose standards and efficient tools, but they might require rethinking, extending, or redesigning and more experimentations.

What's more, further investigations regarding the use of the corpora we developed are still at hand. For instance, a sub-corpus could be extracted from both the OSIAN and MulTed corpora that includes only articles classified based on their topics. This corpus could be considered for use in both training and testing Topic Detection and Tracking (TDT) methods that aim to locate topically related documents in streams of data. Moreover, deeper corpus linguistics analysis is required to investigate Arabic language structure and use. Unfortunately, those investigations were not possible within the timescale of this thesis, and will have to be addressed in the future work.

5. Future Directions

Basically, there are two sources of inspiration for future work:

1) The first one, design, is also the first stage in the procedure of building a corpus. Because without a solid design, everything else is likely to go wrong. So, the focus will be on designing more central and essential corpora that are missing for the Arabic language namely, the Arabic Reference Corpus and Golden Standard Corpus with considerable size and manually annotated with meaningful tags such as lemma, PoS, syntax, and semantic information. During the design process, several critical questions must be answered clearly and in detail. For instance, what we are going to do with these corpora. What research questions need to be answered using those corpora? And what data sources are the most suitable to be used?

   Note that, Arabic Reference and Golden Corpora require the cooperation of many international institutions as well as individual research.

2) The second recommendation for future work is to improve the automatic methods for Arabic text processing and annotation. This can be achieved by investigating hybrid methods and integrating current deep learning technologies that have been shown to have a positive effect on the performance of some NLP and IR systems. Following initial automatic annotation, a manual cross-checking stage is required. For that, online volunteer crowdsourcing including gamification can be used to proofread the automatic morphological and syntactic annotation, and linguist experts will be promoted to a supervisory role which, will save time and reduce the cost of such operations.

6. Closing Remarks

Typically, a resource-poor language refers to the language that lacks "*the basic resources that are fundamental to computational linguistics*" and has a few and small corpora (Zamin et al. 2012). Despite the huge effort made by Arabic corpora builders, Arabic is relatively a resource-poor language when it comes to the availability of free annotated corpora with considerable size and topic variety. This draws our attention, as well as that of other scientific groups, to this field in order to bridge the gap between Arabic and other resource-rich languages such as English, German, and Spanish.

This thesis can be considered as a manual that provides some guidelines for corpora builders interested in developing tools and compiling data to build rich and well-defined Arabic corpora. We attempt to address some of the main challenges faced by Arabic corpora builders and to improve the state-of-the-art for the Arabic language as a whole. Nonetheless, it is clear that we are only at the beginning of establishing reliable Arabic corpus-based studies and many interesting discoveries are yet to be made. Similarly, much work remains to be done for many computational tasks especially across syntax and semantics.

References

Ababneh, Mohamad, Riyad Al-Shalabi, Ghassan Kanaan, and Alaa Al-Nobani. 2012. Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness. *International Arab Journal of Information Technology (IAJIT)* 9(4): 368–372.

Ababou, Nabil., and Azzeddine Mazroui. 2016. A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. *International Journal of Speech Technology (IJST)* 19: 289–302.

Ababou, Nabil, Azzeddine Mazroui, and Rachid Belehbib. 2017. Parsing Arabic Nominal Sentences Using Context Free Grammar and Fundamental Rules of Classical Grammar. *International Journal of Intelligent Systems and Applications (IJISA)* 9(8): 11–24.

Abainia, Kheireddine, Siham Ouamour, and Halim Sayoud. 2016. A novel robust Arabic light stemmer. *Journal of Experimental & Theoretical Artificial Intelligence (JETAI)* 29(3): 557–573.

Abbas, Mourad, Kamel Smaïli, and Daoud Berkani. 2011. Evaluation of Topic Identification Methods on Arabic Corpora. Journal of Digital Information Management (*JDIM)* 9: 185–192.

Abdelali, Ahmed, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)* 14:1044–1054.

Ahmed, Farag, and Andreas Nürnberger. 2008. Arabic/English word translation disambiguation using parallel corpora and matching schemes. In *Proceedings of EAMT* pp.: 8:28.

Aksan, Yesim, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yilmazer, et al. 2012. Construction of the Turkish National Corpus (TNC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)* pp.: 3223–3227.

Al Shamsi, Fatma, and Ahmed Guessoum. 2006. A hidden Markov model-based POS tagger for Arabic. In *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France*, pp.: 31–42.

Al-Bukhitan, Saeed, Tarek Helmy, and Mohammed Al-Mulhem. 2014. Semantic annotation tool for annotating arabic web documents. *Procedia Computer Science* 32: 429–436.

Aldabbas, Omar, Riyad Al-Shalabi, Ghassan Kanan, and Mohammed A. Shehabd. 2016. Arabic Light Stemmer Based on Regular Expression. In *Proceedings of the International Computer Sciences and Informatics Conference (ICSIC 2016)* pp.: 1–9.

Al-Dahdah, A. 1989. The grammar of the Arabic language in tables and lists. *Beirut: Maktabat Lebnan*.

Alfaifi, A. Y. G., Eric Atwell, and I. Hedaya. 2014. Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014* 2: 77–89.

AlGahtani, Shabib, William Black, and John McNaught. 2009. Arabic part-of-speech tagging using transformation-based learning. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools* pp.: 66–70*, Cairo, Egypt*.

Algarni, Mohammed. 2016. Light morphology and Arabic information retrieval. Thesis, University of Canterbury, New Zealand.

Ali, Bilel Ben, and Fethi Jarray. 2013. Genetic approach for Arabic part of speech tagging. *International Journal on Natural Language Computing (IJNLC)* 2(3): 1–12.

Aliwy, Ahmed Hussein. 2013. Arabic Morphosyntactic Raw Text Part of Speech Tagging System. Thesis, University of Warsaw, Poland.

Al-Kabi, Mohammed N., Saif A. Kazakzeh, Belal M. Abu Ata, Saif A. Al-Rababah, and Izzat M. Alsmadi. 2015. A novel root based Arabic stemmer. *Journal of King Saud University-Computer and Information Sciences (JKSU-CIS)* 27: 94–103.

Almujaiwel, Sultan, and Abdulmohsen Al-Thubaity. 2016. Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching. In *Proceedings of the Second International Conference on the Globalization of Second Language Acquisition and Teacher Education (SLATE)*, 2: 103–108, Fukuoka, Japan.

Alotaiby, Fahad, Salah Foda, and Ibrahim Alkharashi. 2010. Clitics in Arabic Language: A Statistical Study. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC)* pp.: 595–601.

Alqrainy, Shihadeh. 2008. A morphological-syntactical analysis approach for Arabic textual tagging. Thesis, De Montfort University, Leicester, UK.

Alrabiah, Maha, A. Al-Salman, and E. S. Atwell. 2013. The design and construction of the 50 million words KSUCCA. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL'2)* pp.: 5–8. The University of Leeds, UK.

Alsaedi, Nasser, Peter Burnap, and Omer Farooq Rana. 2016. Sensing real-world events using Arabic Twitter posts. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)* pp.: 515-518.

Al-Shammari, Eiman, and Jessica Lin. 2008. A novel Arabic lemmatization algorithm. In *Proceedings of the second workshop on Analytics for noisy unstructured text data* pp.: 113–118. Association for Computational Linguistics (ACM).

Al-Shammari, Eiman Tamah. 2013. Lemmatizing, stemming, and query expansion method and system. Google Patents.

Al-Sulaiti, Latifa, and Eric Steven Atwell. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics (IJCL)* 11: 135–171.

Altabba, M., A. Al-Zaraee, and M. A. Shukairy. 2010. An Arabic morphological analyzer and part-of-speech tagger. Thesis, Arab International University, Damascus, Syria.

Al-Thubaity, Abdulmohsen O. 2015. A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation (LREV)* 49: 721–751.

Armstrong, Susan, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky. 2013. *Natural language processing using very large corpora*. Vol. 11. Springer Science & Business Media.

Arts, Tressy, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, and Vit Suchomel. 2014. arTenTen: Arabic Corpus and Word Sketches. *Journal of King Saud University - Computer and Information Sciences (JKSU-CIS)* 26. Special Issue on Arabic NLP: 357–371.

Attia, M., M. Yaseen, and K. Choukri. 2005. *Specifications of the Arabic Written Corpus produced within the NEMLAR project*.

Attia, Mohammed, Ayah Zirikly, and Mona Diab. 2016. The Power of Language Music: Arabic Lemmatization through Patterns. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V),* pp.: 40-50.

Aziz, Wilker, and Lucia Specia. 2011. Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora. In *Proceedings of the 8th Brazilian symposium in information and human language technology (STIL 2011). Cuiabá*, pp.: 234–238.

Ball, Catherine N. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9: 295–302.

Baneyx, Audrey, Jean Charlet, and Marie-Christine Jaulent. 2007. Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. *International Journal of Medical Informatics (IJMI)* 76: 208–215.

Banko, Michele, and Robert C. Moore. 2004. Part of speech tagging in context. In *Proceedings of the 20th international conference on Computational Linguistics*, 556. Association for Computational Linguistics (ACM).

Belinkov, Yonatan, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. 2016. Shamela: A Large-Scale Historical Arabic Corpus. *arXiv preprint arXiv:1612.08989*: 45.

Berners-lee, Tim, James Hendler, and Ora Lassila. 2001. THE SEMANTIC WEB. *Scientific American* 284: 34–43.

Bertels, Ann. 2017. Corpus Linguistics for Language Teaching and LSP. In *Proceedings of the Workshop on Corpus Linguistics*, Universidad de Oriente, Santiago de Cuba.

Bhattacharya, Paheli, Pawan Goyal, and Sudeshna Sarkar. 2016. Query Translation for Cross-Language Information Retrieval using Multilingual Word Clusters. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)* pp.: 152–162, Osaka, Japan.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bodnari, Andreea, Aurélie Névéol, Özlem Uzuner, Pierre Zweigenbaum, and Peter Szolovits. 2013. Multilingual Named-Entity Recognition from Parallel Corpora. In *CLEF (Working Notes)*. Citeseer.

Bojar, Ondrej, Zdenek Zabokrtskỳ, Ondrej Dusek, Petra Galuscáková, Martin Majlis, David Marecek, Jirí Marsík, Michal Novák, Martin Popel, and Ales Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) pp.:* 3921–3928.

Bokova, I. 2012. *World Arabic Language Day*. (*available from http://unesdoc.unesco.org/images/0021/002188/218817e.pdf*)

Bongers, Herman. 1947. *The History and principles of vocabulary control: as it affects in general and of English in particular.* Thesis, Woerden, Holland: Wocopi.

Bouamor, Houda, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) pp.:* 1240–1245.

Boudchiche, Mohamed, and Azzeddine Mazroui. 2016. Approche hybride pour le développement d'un lemmatiseur pour la langue arabe. In *Proceedings of CARI*, 147.

Boudchiche, Mohamed, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2016. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University - Computer and Information Sciences (JKSU-CIS)* 29(2): 141-146.

Boudlal, Abderrahim, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane, MOAO Bebah, and M. Shoul. 2010. Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In *Proceedings of the International Arab conference on information technology*, 1–6. Benghazi Libya.

Boulton, Alex, and Corinne Landure. 2016. Using Corpora in Language Teaching, Learning and Use. *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut* 35.

Brabham, Daren C. 2013. *Crowdsourcing*. Wiley Online Library.

Brants, Thorsten. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, 224–231. Association for Computational Linguistics.

Brants, Thorsten, Wojciech Skut, and Hans Uszkoreit. 2003. Syntactic annotation of a German newspaper corpus. In *Treebanks*, 73–87. Springer.

Brezina, Vaclav, Tony McEnery, and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20: 139–173.

Brits, Karien, Rigardt Pretorius, and Gerhard B. van Huyssteen. 2005. Automatic lemmatization in Setswana: towards a prototype. *South African Journal of African Languages* 25: 37–47.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19: 263–311.

Buckwalter, T. 2004. *Buckwalter Arabic morphological analyzer (BAMA) version 2.0. linguistic data consortium (LDC) catalogue number LDC2004L02*. ISBN1-58563-324-0.

Buckwalter, Tim, and Dilworth Parkinson. 2014. *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Burga, Alicia, Joan Codina, Gabriella Ferraro, Horacio Saggion, and Leo Wanner. 2013. The challenge of syntactic dependency parsing adaptation for the patent domain. In *ESSLLI-13 workshop on extrinsic parse improvement*.

Callan, Jamie, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. *Clueweb09 data set*.

Callison-Burch, Chris, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 175. Association for Computational Linguistics.

Carneiro, Hugo CC, Felipe MG França, and Priscila MV Lima. 2015. Multilingual part-of-speech tagging with weightless neural networks. *Neural Networks* 66: 11–21.

Ćavar, Damir, Alexander Geyken, and Gerald Neumann. 2000. Digital dictionary of the 20th century German language. *Jezikoslovne Tehnologije za Slovenski Jezik: Proceedings of JS*: 110–114.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, 261–268.

Che, Chao, Wenwen Guo, and Jianxin Zhang. 2016. Sentence Alignment Method Based on Maximum Entropy Model Using Anchor Sentences. In *China National Conference on Chinese Computational Linguistics*, 76–85. Springer.

Chen, Yu, and Andreas Eisele. 2012. MultiUN v2: UN Documents with Multilingual Alignments. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) pp.:* 2500–2504.

Chennoufi, Amine, and Azzeddine Mazroui. 2014. Méthodes de lissage d'une approche morpho-statistique pour la voyellation automatique des textes arabes. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, 443–448. Association pour le Traitement Automatique des Langues.

Chennoufi, Amine, and Azzeddine Mazroui. 2016. Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization. *International Journal of Speech Technology* 19: 269–280.

Chouigui, Amina, Oussama Ben Khiroun, and Bilel Elayeb. 2017. ANT Corpus: An Arabic News Text Collection for Textual Classification. In *proceedings of the 14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2017)*, 135–142. Hammamet, Tunisia.

Collins, Peter, and Pam Peters. 1988. The Australian corpus project. *Corpus linguistics, hard and soft*: 103–120.

Cotterell, Ryan, and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)* pp.: 241–245.

Cover, Thomas, and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13: 21–27.

Daelemans, Walter, Antal Van Den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine learning* 34: 11–41.

Daelemans, Walter, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. *arXiv preprint cmp-lg/9607012*.

Darwish, Kareem. 2002. Building a Shallow Morphological Analyzer in One Day. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*.

Deterding, Sebastian, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9–15. ACM.

Diab, Mona. 2007. Towards an optimal POS tag set for Modern Standard Arabic processing. In *Proceedings of recent advances in natural language processing (RANLP)*, 91–96.

Diab, Mona. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools.*

Diab, Mona, Kadri Hacioglu, and Daniel Jurafsky. 2007. Automated methods for processing Arabic text: From tokenization to base phrase chunking. *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer*.

Diab, Mona T. 2007. Improved Arabic base phrase chunking with a new enriched POS tag set. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, 89–96. Association for Computational Linguistics (ACL).

Dror, Judith, Dudu Shaharabani, Rafi Talmon, and Shuly Wintner. 2004. Morphological Analysis of the Qur'an. *Literary and linguistic computing* 19: 431–452.

Dukes, Kais. 2015. Statistical parsing by machine learning from a classical Arabic treebank. *arXiv preprint arXiv:1510.07193*.

Dukes, Kais, and Nizar Habash. 2010. Morphological Annotation of Quranic Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Dušek, Ondrej, Petra Galuščáková, Martin Majliš, David Marecek, Jirı Maršık, Michal Novák, Martin Popel, and Aleš Tamchyna. 2016. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* pp.: 644-648). Association for Computational Linguistics (ACL).

El Hadj, Yahya, I. Al-Sughayeir, and A. Al-Ansari. 2009. Arabic part-of-speech tagging using the sentence structure. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*.

Eldesouki, Mohamed I., Waleed M. Arafa, and K. Darwish. 2009. Stemming techniques of Arabic language: Comparative study from the information retrieval perspective. *The Egyptian Computer Journal (ECJ)* 36(1): 30-49.

El-Haj, Mahmoud, and Rim Koulali. 2013. KALIMAT a multipurpose Arabic Corpus. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2)* pp.: 22–25.

El-Haj, Mahmoud, Udo Kruschwitz, and Chris Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation (LREV)* 49: 549–580.

El-Shishtawy, Tarek, and Fatma El-Ghannam. 2014. A Lemma Based Evaluator for Semitic Language Text Summarization Systems. *arXiv preprint arXiv:1403.5596*.

Fabri, Ray, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner. 2014. Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In *Natural Language Processing of Semitic Languages* pp.: 3–41. Springer.

Farghaly, Ali, and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)* 8(4): 14-33.

Favretti, R. Rossini, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world*: 27–38.

Ferguson, Charles A. 1959. The Arabic Koine. *Language* 35: 616–630.

Field, Anjalie. 2016. An Automated Approach to Syntax-based Analysis of Classical Latin. *Digital Classics Online* 3: 57–78.

Francis, Winthrop Nelson, and Henry Kucera. 1982. Frequency analysis of English usage. Houghton Mifflin.

Freeman, Andrew. 2001. Brill's POS tagger and a Morphology parser for Arabic. In *Proceedings of the ACL/EACL-2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse,France.

Gelbukh, Alexander. 2011. *Computational Linguistics and Intelligent Text Processing*. In *Proceedings of the Seventh International Conference on Computational Linguistics and Natural Language Processing (CICLing).* Tokyo, Japan, Springer.

Germann, Ulrich. 2001. *Aligned hansards of the 36th parliament of canada*. Release 2001-1a. *(available from https://www.isi.edu/natural-language/download/hansard/)*

Gharaibeh, Islah K., and Natheer K. Gharaibeh. 2012. Towards Arabic noun phrase extractor (ANPE) using information retrieval techniques. *Software Engineering* 2: 36–42.

Giménez, Jesús, and Lluis Marquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)* pp.: 759–765.

Goldhahn, Dirk, Maciej Sumalvico, and Uwe Quasthoff. 2016. Corpus Collection for Under- Resourced Languages with More than One Million Speakers. In *Collaboration and Computing for Under-Resourced Languages (CCURL)*, 67–73. Portorož.

Goutte, Cyril, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego California, USA.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 439–448. ACM.

Habash, Nizar, Mona T. Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)* pp.: 711–718.

Habash, Nizar, and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* pp.: 573–580. Association for Computational Linguistics (ACL).

Habash, Nizar, Owen Rambow, and Ryan Roth. 2009. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)* pp.: 102–109*, Cairo, Egypt*.

Habash, Nizar Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3: 1–187.

Habernal, Ivan, and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hadni, Meryeme, Said Alaoui Ouatik, Abdelmonaime Lachkar, and Mohammed Meknassi. 2013. Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. *International Journal on Natural Language Computing* 2: 1–15.

Hajic, Jan, Eva Hajicová, Jarmila Panevová, Petr Sgall, Ondrej Bojar, Silvie Cinková, Eva Fucíková, et al. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) pp.:* 3153–3160.

Halliday, Michael, Christian MIM Matthiessen, and Christian Matthiessen. 2014. *An introduction to functional grammar*. Routledge.

Hammouda, Faten Khalfallah, and Abdelsalam Abdelhamid Almarimi. 2010. Heuristic Lemmatization for Arabic Texts Indexation and Classification. *Journal of Computer Science (JCS)* 6 (6): 660-665.

Harrat, Salima, Karima Meftouh, and Kamel Smaili. 2017. Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*.

Hu, Kaibao. 2016. Corpus-Based Translation Studies: Problems and Prospects. In *Introducing Corpus-based Translation Studies*, pp.: 223–233. Springer.

Hu, Kaibao, and others. 2016. *Introducing corpus-based translation studies*. Book Edition, New Frontiers in Translation Studies, Springer.

Hunston, Susan. 2013. Corpus Linguistics: Historical Development. *The Encyclopedia of Applied Linguistics*.

Zeroual, Imad, and Abdelhak Lakhouaja. 2016. Adapting a decision tree based tagger for Arabic. In *Proceedings of the Information Technology for Organizations Development (IT4OD)* pp.: 1–6. IEEE.

Ishkewy, Hossam, Hany Harb, and Hassan Farahat. 2014. Azhary: An Arabic Lexical Ontology. *International Journal of Web& Semantic Technology (IJWesT)* 5(4): 71:82.

Jaafar, Younes, Driss Namly, Karim Bouzoubaa, and Abdellah Yousfi. 2016. Enhancing Arabic stemming process using resources and benchmarking tools. *Journal of King Saud University-Computer and Information Sciences (JKSU-CIS)* 29(2): 164-170.

Jacob, Alen, Amal Babu, and PC Reghu Raj. 2015. TnT tagger with fuzzy rule based learning. In Proceedings of the *Signal Processing, Informatics, Communication and Energy Systems (SPICES)* pp.: 1–5. IEEE.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *Proceedings of the Seventh International Corpus Linguistics Conference* pp.: 125–127. Lancaster, UK.

James, Orr. 2015. *The International Standard Bible Encyclopedia*. Delmarva Publications, Inc.

Järvinen, Timo. 1994. Annotating 200 million words: the Bank of English project. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, 565–568. Association for Computational Linguistics (ACL).

Joachims, Thorsten. 1999. Making large-scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning. B.Schlkopf and C. Burges and A. Smola (ed.), MIT Press.

de Jong, F. M. G., Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)* pp.: 3259–3264.

Jrai, Enas Mahmoud Abu. 2013. Hybrid Technique for Arabic Text Compression. Ms Dissertation, Middle East University, Jordan.

Kadim, Ayoub, and Azzeddine Lazrek. 2016. Bidirectional HMM-based Arabic POS tagging. *International Journal of Speech Technology (IJST)* 19: 303–312.

Kennedy, Graeme. 2014. *An introduction to corpus linguistics*. Book Edition, Routledge.

Khalifa, Salam, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A Large Scale Corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960*.

Khalifa, Salam, Sara Hassan, and Nizar Habash. 2017. A Morphological Analyzer for Gulf Arabic Verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop (co-located with EACL 2017)* pp.: 35-45.

Khoja, Shereen. 2001. APT: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at NAACL*, 20–25.

Khoja, Shereen, and Roger Garside. 1999. Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

Khorsheed, Mohammed S., Khaled M. Alhazmi, and Adil M. Asiri. 2009. Developing typewritten Arabic corpus with multi-fonts (TRACOM). In *Proceedings of the International Workshop on Multilingual OCR*, 16. ACM.

Kilgarriff, Adam. 2013. Using corpora as data source for dictionaries. *The Bloomsbury Companion to Lexicography. London: Bloomsbury*: 77–96.

Kim, Hansaem. 2006. Korean national corpus in the 21st century sejong project. In *Proceedings of the 13th NIJL International Symposium*, 49–54. National Institute for Japanese Language Tokyo.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, 5:79–86. Citeseer.

Kotelnikov, Evgeny, Elena Razova, and Irina Fishcheva. 2017. A Close Look at Russian Morphological Parsers: Which One Is the Best? In *Proceedings of the Conference on Artificial Intelligence and Natural Language* pp:.131–142. Springer.

Kovář, Vojtěch, Vít Baisa, and Miloš Jakubíček. 2016. Sketch Engine for bilingual lexicography. *International Journal of Lexicography (IJL)* 29: 339–352.

Kreaa, Abdel Hamid, Ahmad S. Ahmad, and Kassem Kabalan. 2014. Arabic words stemming approach using Arabic WordNet. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 4: 1.

Kübler, Sandra, and Emad Mohamed. 2012. Part of speech tagging for Arabic. *Natural Language Engineering* (NLE) 18: 521–548.

Lamraoui, Fethi, and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*.

Larkey, Leah S., Lisa Ballesteros, and Margaret E. Connell. 2007. Light stemming for Arabic information retrieval. In *Arabic computational morphology*, pp.: 221–243. Springer, Dordrecht.

Le, Hoang Thi My, Phan Thi Bong, and Phan Huy Khanh. 2012. Building a machine translation system in a restrict context from Ka-Tu Language into Vietnamese. In *Proceedings of the Fourth International Conference on Knowledge and Systems Engineering (KSE)* pp.: 167–172. IEEE.

Leech, G. N. 1991. *The state of the art in corpus linguistics. English Corpus Linguistics: Studies in Honor of Jan Svartuk. Eds. K. Aijmer & B. Altenberg*. London: Longman.

Leech, Geoffrey. 1992a. Corpora and theories of linguistic performance. *Directions in corpus linguistics*: 105–122.

Leech, Geoffrey. 1992b. 100 million words of English: the British National Corpus (BNC). *Language Research* 28: 1–13.

Lefever, Els, and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) 2*: 158–166.

Lefever, Els, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologie 2*: 317–322. Association for Computational Linguistics (ACL).

Li, Lei, Corina Forascu, Mahmoud El-Haj, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian. In Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization pp.: 1–12, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

Liua, Quan, Hui Jiangb, Zhen-Hua Linga, Xiaodan Zhuc, Si Weid, and Yu Hua. 2016. Commonsense Knowledge Enhanced Embeddings for Solving Pronoun Disambiguation Problems in Winograd Schema Challenge. *arXiv preprint arXiv:1611.04146*.

Lüdeling, Anke, and Merja Kytö. 2008. *Corpus linguistics: an international handbook*. Vol. 1. 2 vols. Walter de Gruyter.

Lund, Lisa, and Patrick O'Regan. 2016. *Gamifying Natural Language Acquisition: A quantitative study on Swedish antonyms while examining the effects of consensus driven rewards*. Degree Project, KTH Royal Institute of Technology, Stockholm, Sweden.

Maabid, Abdelmawgoud Mohamed, Tarek Elghazaly, and Mervat Ghaith. 2015. An Enhanced Rule Based Arabic Morphological Analyzer Based on Proposed Assessment Criteria. In *Advances in Swarm and Computational Intelligence*, ed. Ying Tan, Yuhui Shi, Fernando Buarque, Alexander Gelbukh, Swagatam Das, and Andries Engelbrecht, 393–400. Lecture Notes in Computer Science 9142. Springer International Publishing.

Maamouri, Mohamed, and Ann Bies. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages* pp.: 2–9. Association for Computational Linguistics (ACL).

Maamouri, Mohamed, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)* pp.: 2348–2354.

Maamouri, Mohamed, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri, and Wadji Zaghouani. 2013. Arabic Treebank: Part 1 v 4.1.

Maegaard, Bente, Mohammed Attia, Khalid Choukri, Steven Krauwer, Chafik Mokbel, and Mustafa Yaseen. 2009. MEDAR: Arabic language technology, state-of-the-art and a cooperation roadmap. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. Citeseer.

Magdy, Walid, and Gareth JF Jones. 2014. Studying machine translation technologies for large-data CLIR tasks: a patent prior-art search case study. *Information retrieval* 17: 492–519.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Vol. 999. MIT Press.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pp.: 55–60.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19: 313–330.

Marques, Nuno C., and Gabriel Pereira Lopes. 1996. Using neural nets for portuguese part-of-speech tagging. In *Proceedings of the Fifth International Conference on the Cognitive Science of Natural Language Processing (CSNLP)*. Citeseer.

McEnery, Anthony M., and Anita Wilson. 2001. *Corpus linguistics: an introduction.* Edinburgh University Press.

McEnery, T., and A. Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony, and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mdhaffar, Salima, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP) (co-located with EACL 2017)* pp.: 55-61.

Milfull, Inge. 2009. Mutual Illumination: The Dictionary of Old English and the Ongoing Revision of the Oxford English Dictionary (OED3). *Florilegium* 26: 235–264.

Mohammadi, Mehdi, and Nasser GhasemAghaee. 2010. Building bilingual parallel corpora based on wikipedia. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, 2:264–268. IEEE.

Moore, Robert C. 2002. *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Mostefa, Djamel, Mariama Laïb, Stéphane Chaudiron, Khalid Choukri, and G. Chalendar. 2009. A multilingual named entity corpus for Arabic, English and French. *MEDAR* 2009: 2nd.

Mustafa, Suleiman H. 2012. Word stemming for Arabic information retrieval: The case for simple light stemming. *Abhath Al-Yarmouk: Science & Engineering Series* 21: 2012.

Nakov, Preslav. 2014. Web as a Corpus: Going Beyond the n-gram. In *Russian Summer School in Information Retrieval*, 185–228. Springer.

Naseem, Tahira, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2014. Multilingual Part-of-Speech Tagging: Two Unsupervised Approaches. *arXiv preprint arXiv:1401.5695*.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring natural language: working with the British component of the International Corpus of English*. Vol. 29. John Benjamins Publishing.

Neme, Alexis Amid, and Eric Laporte. 2013. Pattern-and-root inflectional morphology: the Arabic broken plural. *Language Sciences* 40: 221–250.

Neuhoff, David L. 1975. The Viterbi algorithm as an aid in text recognition (Corresp.). *Information Theory, IEEE Transactions on* 21: 222–226.

Neves, Mariana, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Paris, France. European Language Resources Association (ELRA)*.

Ney, Hermann, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language* 8: 1–38.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194: 151–175.

Östling, Robert, and Jörg Tiedemann. 2016. Efficient Word Alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics* 106: 125.

Otair, M. 2013. Comparative analysis of Arabic stemming algorithms. *International Journal of Managing Information Technology (IJMIT) Vol* 5: 1–12.

Parker, and Robert. 2009. *Arabic Gigaword Fourth Edition LDC2009T30*.

Parkinson, Dilworth B. 2012. *ArabiCorpus*.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth. 2014a. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland*.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014b. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)* 14:1094–1101.

Paul, Michael, Marcello Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 evaluation campaign. In *IWSLT*, 10:3–27.

Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Quasthoff, Uwe, Dirk Goldhahn, and Thomas Eckart. 2014. Building large resources for text mining: The Leipzig Corpora Collection. In *Text Mining*, 3–24. Springer.

Rabiee, Hajder S. 2011. Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic. In *RANLP Student Research Workshop*, 127–132.

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77: 257–286.

Rambow, Owen, Bonnie Dorr, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, et al. 2006. Parallel syntactic annotation of multiple languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genoa, Italy*.

Ratnaparkhi, Adwait, and others. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, 1:133–142. Philadelphia, USA.

Regragui, Yassir, Lahsen Abouenour, Fettoum Krieche, Karim Bouzoubaa, and Paolo Rosso. 2016. Arabic WordNet: New Content and New Applications. In *Proceedings of the Eighth Global WordNet Conference* pp.: 330-338.

Rissanen, Matti, and others. 1993. The Helsinki Corpus of English Texts. Book edition: *Corpora across the centuries: proceedings of the First International Colloquium on English Diachronic Corpora 11*: 73–81, St Catharine's College Cambridge, UK.

Rogati, Monica, Scott McCarley, and Yiming Yang. 2003. Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, 391–398. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL).

Rozovskaya, Alla, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The Second QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing (ANLP)* pp.: 26-35.

Saad, E. - S.M, and et al. 2005. *An Improved Arabic morphology analyzer (IAMA)*.

Saad, Motaz K., and Wesam Ashour. 2010. OSAC: Open Source Arabic Corpora. In *Proceedings of the sixth 6th International Symposium on Electrical and Electronics Engineering and Computer Science (EEECS*) 10: 118-123, European University of Lefke, Cyprus.

Sakho, Mohamed Lemine. 2012. Teaching Arabic as a Second Language in International School in Dubai A case study exploring new perspectives in learning materials design and development. Ms Dissertation, British University in Dubai, UAE.

Salloum, Wael, and Nizar Habash. 2014. Adam: Analyzer for dialectal Arabic morphology. *Journal of King Saud University-Computer and Information Sciences (JKSU-CIS)* 26: 372–378.

Samih, Younes, Mohammed Attia, Mohamed Eldesouki, Hamdy Mubarak, Ahmed Abdelali, Laura Kallmeyer, and Kareem Darwish. 2017. A Neural Architecture for Dialectal Arabic Segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP) (co-located with EACL 2017)* pp.: 46-54.

Sawalha, Majdi. 2009. Arabic Morphological Features Tag set. University of Leeds.

Sawalha, Majdi, Claire Brierley, and Eric Atwell. 2014. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning (version 2.0). In *Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop 31st May 2014, Reykjavik, Iceland*, 42–47. The University of Leeds.

Scherrer, Yves, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. SwissAdmin: A multilingual tagged parallel corpus of press releases.

Schmid, Helmut. 1994a. Part-of-speech tagging with neural networks. In *Proceedings of the 15th Conference on Computational Linguistics,* 1: 172–176. Association for Computational Linguistics (ACL).

Schmid, Helmut. 1994b. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 12:44–49. Citeseer.

Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pp.: 13–25. Springer.

Schmid, Helmut. 2013. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, 154. Routledge.

Sharaf, Abdul-Baquee M., and Eric Atwell. 2012a. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC*).

Sharaf, Abdul-Baquee M., and Eric Atwell. 2012b. QurSim: A corpus for evaluation of relatedness in short texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC*) pp.: 2295–2302.

Shastri, Srikant V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME journal* 12.

Sinclair, John. 1996. Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P (available from http://www.ilc.pi.cnr.it/EAGLES/corpustyp/corpustyp.html)*.

Sinclair, John. 2004. Intuition and annotation–the discussion continues. *Language and Computers* 49: 39–59.

Sinclair, John. 2005. Corpus and text-basic principles. *Developing linguistic corpora: A guide to good practice*, 92: 1–16. Oxford: Oxbow Books.

Sinclair, John. 2008. Borrowed ideas. *Language and computers studies in practical linguistics* 64: 21.

Sinclair, John. 1987. *Collins COBUILD English language dictionary*. Harper Collins Publishers.

Singh, Jasmeet, and Vishal Gupta. 2016a. Text Stemming: Approaches, Applications, and Challenges. *ACM Computing Surveys (CSUR)* pp.: 49: 45.

Singh, Jasmeet, and Vishal Gupta. 2016b. A systematic review of text stemming techniques. *Artificial Intelligence Review* pp.: 1–61.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Stolcke, Andreas, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*. Vol. 5.

Suchomel, Vít, Jan Pomikálek, and others. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pp.: 39–43.

Swuaileh, Reem, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. ArabicWeb16: A New Crawl for Today's Arabic Web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 673–676. ACM.

Sylak-Glassman, John, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology*, 72–93. Springer.

Taghva, Kazem, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC*) 1:152–157. IEEE.

Teubert, Wolfgang. 2015. Corpus linguistics and lexicography: The beginning of a beautiful friendship. *Issues* 31.

Tiam-Lee, Thomas James, and Solomon See. 2014. Building a sentiment corpus using a gamified framework. In *Proceedings of the International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM),* pp.: 1–8. IEEE.

Tian, Liang, Fai Wong, and Sam Chao. 2011. Word alignment using GIZA++ on Windows. *Machine Translation*.

Tiedemann, Jörg. 2007. Building a multilingual parallel subtitle corpus. In *Proceedings of the 18th Meeting of Computational Linguistics in the Netherlands (CLIN)*.

Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) pp.:* 2214–2218.

Tiedemann, Jörg, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Ostling, and M. Weller-Di Marco. 2016. Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany.* Association for Computational Linguistics (ACL).

Tlili-Guiassa, Yamina. 2006. Hybrid method for tagging Arabic text. *Journal of Computer Science (JCS)* 2: 245–248.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 173–180. Association for Computational Linguistics (ACL).

Toutanova, Kristina, and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics,* 13: 63–70. Association for Computational Linguistics (ACL).

Tsarfaty, Reut, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics* 39: 15–22.

Utiyama, Masao, and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. *Proceedings of MT summit XI*: 475–482.

Utvić, Miloš. 2011. Annotating the corpus of contemporary Serbian. In *Proceedings of the INFOtheca '12 Conference*.

Van den Bosch, Antal, Erwin Marsi, and Abdelhadi Soudi. 2007. Memory-based morphological analysis and part-of-speech tagging of Arabic. In *Arabic Computational Morphology*, 201–217. Springer.

Vandeghinste, Vincent, and Erik F. Tjong Kim Sang. 2004. Using a Parallel Transcript/Subtitle Corpus for Sentence Compression. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Vapnik, Vladimir. 2013. *The nature of statistical learning theory*. Springer Science & Business Media.

Váradi, Tamás. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC*).
Versteegh, Kees. 2014. *The Arabic language*. Book Edition. Edinburgh University Press.
Vladimir, Vapnik N., and V. Vapnik. 1995. *The nature of statistical learning theory*. Springer Heidelberg.
Watson, Janet CE. 2002. *The phonology and morphology of Arabic*. Oxford University Press on Demand.
van der Wees, Marlies, Arianna Bisazza, and Christof Monz. 2016. Measuring the Effect of Conversational Aspects on Machine Translation Quality. In Proceedings of the 26th International Conference on Computational Linguistics (COLING): Technical Papers, pp.: 2571-2581.
Wehrli, Eric, and Luka Nerima. 2015. The Fips multilingual parser. In *Language Production, Cognition, and the Lexicon*, pp.: 473–490. Springer.
Wilson, D. Randall. 1997. Advances in instance-based learning algorithms. Dissertation, Brigham Young University. USA.
Xing, Junwen, Derek F. Wong, Lidia S. Chao, Ana Luisa V. Leal, Marcia Schmaltz, and Chunhui Lu. 2016. Syntaxtree aligner: A web-based parallel tree alignment toolkit. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)* pp.: 37–42. IEEE.
Yang, Christopher C., and Kar Wing Li. 2004. Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Information processing & management* 40: 939–955.
Yaseen, Mustafa, M. Attia, Bente Maegaard, Khalid Choukri, N. Paulsson, S. Haamid, Steven Krauwer, et al. 2006. Building annotated written and spoken Arabic LR's in NEMLAR project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC) pp.:* 533–538.
Yassein, Muneer Bani, and Yarub A. Wahsheh. 2016. HQTP v. 2: Holy Quran Transfer Protocol version 2. In *Proceedings of the Seventh International Conference on Computer Science and Information Technology (CSIT)*, pp.: 1–5. IEEE.
Zaghouani, Wajdi. 2014. Critical survey of the freely available Arabic corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), OSACT Workshop. Rejkavik, Iceland*.
Zaghouani, Wajdi, Nizar Habash, and Behrang Mohit. 2014. *The qatar arabic language bank guidelines*. Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September.
Zamin, Norshuhani, Alan Oxley, Zainab Abu Bakar, and Syed Ahmad Farhan. 2012. A statistical dictionary-based word alignment algorithm: An unsupervised approach. In *Proceedings of the International Conference on Computer & Information Science (ICCIS)* 1:396–402. IEEE.
Zeroual, Imad, and Abdelhak Lakhouaja. 2014. Clitiques-Stemmer: nouveau stemmer pour la langue Arabe. In *Proceedings of the First National Doctoral Symposium on Arabic Language Engineering (JDILA)*, Rabat, Morocco.
Zeroual, Imad, Mohamed Boudchiche, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2015. Improving Arabic light stemming algorithm using linguistic resources. In *Proceedings of the First National Doctoral Symposium on Arabic Language Engineering (JDILA),* Fez, Morocco.
Zeroual, Imad, Anoual El Kah, and Abdelhak Lakhouaja. 2017. Gamification for Arabic Natural Language Processing: Ideas into Practice. *Transactions on Machine Learning and Artificial Intelligence (TMAI)* 5(4).
Zeroual, Imad, and Abdelhak Lakhouaja. 2016. A new Quranic Corpus rich in morphosyntactical information. *International Journal of Speech Technology* (IJST) 19(2): 339-346.
Zeroual, Imad, and Abdelhak Lakhouaja. 2018. Data science in light of natural language processing: An overview. *Procedia Computer Science* 127: 82–91.
Zeroual, Imad, Abdelhak Lakhouaja, and Rachid Belahbib. 2017. Towards a standard Part of Speech tagset for the Arabic language. *Journal of King Saud University - Computer and Information Sciences* (JKSU-CIS) 29: 174–181.
Zerrouki, Taha, and Amar Balla. 2017. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief* 11: 147–151.

Ziemski, Michal, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
Zitouni, Imed, ed. 2014. *Natural Language Processing of Semitic Languages*. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer Berlin Heidelberg.

Appendix A: Al Mus'haf Corpus

الْحَمْدُ
Stem:حمد/Hmd|StemPat:فَعْلُ/faEolu|POS:مأ/VN|Lem:حَمْد/Hamod|LemPat:فَعْل/faEol|Root:حمد/Hmd
لِلَّهِ
Stem:لله/llh|StemPat:#|POS:اسجل/MN|Lem:لله/llh|LemPat:#|Root:#
رَبِّ
Stem:رب/rb|StemPat:فَعْلِ/faEoli|POS:اسج/NN|Lem:رَبّ/rab~|LemPat:فَعْل/faEol|Root:ربب/rbb
الْعَالَمِينَ
Stem:عالمين/Ealmyn|StemPat:فَاعَلِينَ/faAEaliyna|POS:اسج/NN|Lem:عَالَم/EaAlam|LemPat:فَاعَل/faAEal|Root:علم/Elm

Figure 1: A sample of Al Mus'haf corpus encoded in TXT format (one word per line)

Words	Stem	Stem Pattern	POS tag	Lemma	Lemma Pattern	Root
الْحَمْدُ AloHamodu	حمد Hmd	فَعْلُ faEolu	مأ VN	حَمْد Hamod	فَعْل faEol	حمد Hmd
لِلَّهِ lil~ahi	لله llh	#	اسجل MN	لله Llh	#	#
رَبِّ rab~i	رب rb	فَعْلِ faEoli	اسج NN	رَبّ ~rab	فَعْل faEol	ربب rbb
الْعَالَمِينَ AloEaAlamiyna	عالمين Ealmyn	فَاعَلِينَ faAEaliyna	اسج NN	عَالَم EaAlam	فَاعَل faAEal	علم Elm

Figure 2: A sample of Al Mus'haf corpus encoded in CSV format

Appendix B: Survey

Table 1: Corpora of the survey

Corpora	language(s)	Corpus reference (URL or DOI)
Al-Hayat Arabic Corpus	Arabic	http://catalog.elra.info/product_info.php?products_id=632
Alpino Treebank	Dutch	http://odur.let.rug.nl/~vannoord/trees/
Amara	20 languages	http://alt.qcri.org/resources/qedcorpus/
Arab-Acquis	Arabic, English, and French	http://www.aclweb.org/anthology/E/E17/E17-2.pdf#page=267
Arabic English Parallel News	Arabic, English	https://catalog.ldc.upenn.edu/LDC2004T18
Arabic Treebank	Arabic	https://catalog.ldc.upenn.edu/LDC2005T20
Aranea Web Corpora	18 languages	http://unesco.uniba.sk/guest/
ARCADE/ROMANSEVAL	English, French, Italian	http://catalog.elra.info/product_info.php?products_id=535
ARCHER	English	http://www.projects.alc.manchester.ac.uk/archer/
arTenTen	Arabic	https://www.sketchengine.co.uk/artenten-corpus/
BAF	French, English	http://rali.iro.umontreal.ca/rali/?q=fr/BAF
BoLC	Italian/English	http://corpora.ficlit.unibo.it/
BOLT	English, Chinese	https://catalog.ldc.upenn.edu/LDC2016T19
Brown	English	http://clu.uni.no/icame/brown/bcm.html
BulTreeBank	Bulgarian	http://www.bultreebank.org/
CELEX2	English, German, Dutch	https://catalog.ldc.upenn.edu/LDC96L14
CELT	Irish, Latin, English, French, Spanish, Italian, Provençal, Dutch, Danish	http://www.ucc.ie/celt/
CETEMPúblico	Portuguese	http://www.linguateca.pt/CETEMPublico/
CINTIL Corpus	Portuguese	http://cintil.ul.pt/
ConVote	English	http://www.cs.cornell.edu/home/llee/data/convote.html
CORGA	Galician	http://corpus.cirp.es/corga/
CORIS	Italian	http://corpora.ficlit.unibo.it/
CORIS/CODIS	Italian	http://corpora.ficlit.unibo.it/
Corpora for eContent professionals	Greek-English, Bulgarian-English, Slovene-English, and Serbian-English	http://dl.acm.org/citation.cfm?id=1706253&CFID=770410841&CFTOKEN=72713156
Corpus "TUITS" IRÓNICOS	Spanish	https://ivanvladimir.github.io/sitio-corpus-ironia/
Corpus del Español	Spanish	http://www.corpusdelespanol.org/hist-gen/
Corpus del Español (Web)	Spanish	http://www.corpusdelespanol.org/web-dial/
Corpus do Português	Portuguese	http://www.corpusdoportugues.org/hist-gen/2008/
Corpus do Português (Web)	Portuguese	http://www.corpusdoportugues.org/web-dial/
Corpus of Spoken Lithuanian	Lithuanian	http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas/
CRATER	English, French, Spanish	http://catalog.elra.info/product_info.php?products_id=636
CREA	Spanish	corpus.rae.es/creanet.html
Croatian National Corpus	Croatian	10.1007/978-1-4020-4068-9_14
Daniel corpus	Chinese, English, Greek, Polish, Russian	10.13140/2.1.1094.6881
DeReKo	German	http://www1.ids-mannheim.de/kl/projekte/dereko_i.html
deWaC	German	http://wacky.sslmit.unibo.it
DiaCORIS	Italian	http://corpora.ficlit.unibo.it/
EMEA Corpus	22 languages	http://opus.lingfil.uu.se/EMEA.php
English Gigaword	English	https://catalog.ldc.upenn.edu/ldc2011t07
Europarl	21 European languages	http://www.statmt.org/europarl/
Frantext	French	http://www.frantext.fr/
frWaC	French	http://wacky.sslmit.unibo.it
GeFRePaC	German, and French	http://catalog.elra.info/product_info.php?products_id=633
GENIA	English	http://www.nactem.ac.uk/meta-knowledge/download.php
Global English Monitor Corpus	English	http://www.corpus.bham.ac.uk/ccl/global.htm
GUM corpus	English	https://corpling.uis.georgetown.edu/gum/
Helsinki Corpus	English	http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus
ICE-GB	English	http://www.ucl.ac.uk/english-usage/projects/ice-gb/
International Corpus of English	English	http://www.ucl.ac.uk/english-usage/projects/ice.htm
International Corpus of Learner English - ICLE	25 languages	http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm
INTERSECT	English, French, German	http://arts.brighton.ac.uk/staff/raf-salkie/intersect
itWaC	Italian	http://wacky.sslmit.unibo.it
KACST	Arabic	10.1007/s10579-014-9284-1
KAzakh Dependency Treebank	Kazakh	https://github.com/UniversalDependencies/UD_Kazakh
Kazakh Language Corpus	Kazakh	http://kazcorpus.kz/klcweb/en/
Korean National Corpus	Korean	http://www.sejong.or.kr/

KorpusDK	Danish	http://ordnet.dk/korpusdk_en?set_language=en
KSUCCA)	Arabic	http://ksucorpus.ksu.edu.sa
Lancaster Parsed Corpus	English	http://clu.uni.no/icame/lanpeks.html
Lancaster-Leeds Treebank	English	http://universal.elra.info/product_info.php?cPath=42_43&products_id=437
LASSY	Dutch	http://odur.let.rug.nl/~vannoord/Lassy/
LIVAC	Mandarin Chinese	http://www.livac.org
Malay Concordance Project	Malay	http://mcp.anu.edu.au/
Movie Review Data	English	http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/
MultiLing Multilingual Multi-Document Summarization Corpus	Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish	http://multiling.iit.demokritos.gr/pages/view/1540/task-mms-multi-document-summarization-data-and-information
MultiUN	English, French, Spanish, Arabic, Russian, Chinese, German	http://www.euromatrixplus.net/multi-un/
Negra	German	http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
NEMLAR Corpus	Arabic	http://www.rdi-eg.com/Projects/nemlar.htm
OntoNotes	English, Chinese, Arabic	https://catalog.ldc.upenn.edu/LDC2013T19
PANACEA	English, French, Greek	http://catalog.elra.info/product_info.php?products_id=1182
Parallela	English-Polish	http://paralela.clarin-pl.eu/
ParTUT	English, French, Italian	https://github.com/msang/partut-repo
Penn Treebank	English	https://catalog.ldc.upenn.edu/ldc99t42
Polarity	English	10.1109/HNICEM.2014.7016215
PTPARL Corpus	Portuguese	http://catalog.elra.info/product_info.php?products_id=1179
Russian collection	Russian	http://corpus.leeds.ac.uk/ruscorpora.html
Russian National Corpus	Russian	http://www.ruscorpora.ru
SIKOR	North Saami, South Saami, Aanaar Saami, Lule Saami, Skolt Saami	http://gtweb.uit.no/korp/#?cqp=%5B%5D&lang=en
Sinica	Chinese	http://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm
Slovak National Corpus	Slovak	http://korpus.juls.savba.sk/
Syntactic Database for modern Spanish (BDS)	Spanish	http://www.bds.usc.es/
The American National Corpus	English	http://www.anc.org/
The Bank of English	English	http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html
The British National Corpus	English	http://www.natcorp.ox.ac.uk/
The Czech National Corpus	Czech	https://www.korpus.cz/
The Hellenic National Corpus	Greek	http://hnc.ilsp.gr/find.asp
The hungarian gigaword corpus	Hungarian	http://corpus.nytud.hu/mnsz/index_eng.html
The International Corpus of Arabic	Arabic	http://www.bibalex.org/ica
The Lancaster Corpus of Mandarin Chinese	Chinese	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/default.htm
The National Corpus of Polish	Polish	http://nkjp.pl
The New York Times Annotated Corpus	English	https://catalog.ldc.upenn.edu/LDC2008T19
TIGER Corpus	German	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
TIPSTER Complete	English	https://catalog.ldc.upenn.edu/LDC93T3A
TüBa-D/Z treebank	German	http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html
Turkish National Corpus	Turkish	http://www.tnc.org.tr
UKPConvArg	English	https://www.ukp.tu-darmstadt.de/data/argumentation-mining/ukpconvarg1-corpus/
ukWaC	English	http://wacky.sslmit.unibo.it
Wikipedia: Database	More than 270 languages	https://en.wikipedia.org/wiki/Wikipedia:Database_download
WIT[3]	109 languages	https://wit3.fbk.eu/
WOCHAT	English	http://workshop.colips.org/wochat/

Appendix C: OSIAN Corpus

Table 1: List of crawled web-domains

Region or country	Web-domain	Nb. of articles
International	news.un.org arabic.euronews.com ara.reuters.com namnewsnetwork.org arabic.sputniknews.com	693,629
Middle-east	aljazeera.net alarabiya.net	366,211
Algeria	djazairess.com	588,514
Australia	eltelegraph.com	4,614
Canada	arabnews24.ca halacanada.ca	30,135
China	arabic.cctv.com	1,365
Egypt	alwatanalarabi.com	85,351
France	france24.com	74,718
Iran	alalam.ir	344,011
Iraq	iraqakhbar.com	28,248
Germany	dw.com	117,261
Jordon	sarayanews.com	49,461
Morocco	www.marocpress.com	188,045
Palestine	al-ayyam.ps	81,495
Qatar	raya.com	8,986
Russia	arabic.rt.com	57,238
Saudi Arabia	alwatan.com.sa	1,512
Sweden	alkompis.se	33,790
Syria	syria.news	36542
Tunisia	www.turess.com	495,674
Turkey	turkey-post.net aa.com.tr	76,638
UAE	emaratalyoum.com	25,081
UK	bbc.com	10,686
USA	arabic.cnn.com	113,557

Appendix D: MulTed Corpus

Table 1: The number of segments pairs with English of the second top 15 languages in MulTed corpus

Languages	Nb. of files	Nb. of tokens	Nb. of segments
Polish	932	1,297,125	228,738
Greek	898	1,582,023	222,279
Deutsch	894	1,604,856	226,573
Turkish	885	1,239,183	224,495
Serbian	871	1,378,420	212,990
Hungarian	800	1,188,401	202,221
Bulgarian	791	1,425,035	205,132
Portuguese	790	1,435,525	199,333
Farsi	787	1,673,768	193,726
Ukrainian	680	1,016,555	164,927
Croatian	648	1,019,459	158,525
Thai	624	270,437	145,432
Czech	581	880,445	139,670
Indonesian	509	764,600	113,990
Slovenian	494	607,375	96,943

```
<?xml version="1.0" encoding="UTF-8"?>
 <Talk id="Fg_JcKSHUtQ">
  <Category>Architecture and Design</Category >
  <Title>A robot that flies like a bird</Title>
  <Speaker>Markus Fischer</Speaker>
  <Time-slot>00:00:15,259 --> 00:00:18,259</Time-slot>
    <Segment id="1">
      <Word PoS="PRON" Lemma="It">It</Word>
      <Word PoS="VERB" Lemma="be">is</Word>
      <Word PoS="DET" Lemma="a">a</Word>
      <Word PoS="NOUN" Lemma="dream">dream</Word>
      <Word PoS="ADP" Lemma="of">of</Word>
      <Word PoS="NOUN" Lemma="mankind">mankind</Word>
      <Word PoS="SENT" Lemma="unknown">,</Word>
    </Segment>
    …
 </Talk>
```

Figure 1: A sample of a PoS tagged version of an English subtitle