Accepted in Colloque pour les Etudiants Chercheurs en Traitement Automatique du Langage Naturel et ses applications (CEC-TAL 2015). Sousse - Tunisie, le 23-25 Mars 2015.

Development of Arabic particles lexicon using the LMF framework

Driss Namly *, Karim Bouzoubaa *, Youssef Tahir **, Hakima Khamar ***
*Ecole mohammadia d'ingénieurs, Université mohammed V Rabat, Maroc
**Ecole Nationale Supérieure d'Arts et Métiers, Casablanca
*** Faculté des lettres et des sciences humaines, Université mohammed V Rabat,
namly_driss@yahoo.fr; karim.bouzoubaa@emi.ac.ma; ytahirensamcasa@gmail.com; khamar@um5a.ac.ma

RÉSUMÉ. Le développement technologique et la croissance rapide dans le domaine du Traitement Automatique de la Langues Arabe (TALA) offrent des applications de plus en plus performantes et engendrent un intérêt particulier pour les ressources linguistiques. Cet article traite le développement de la ressource des particules Arabes. Ce choix s'explique par l'importance des particules dans la constitution d'une phrase. A titre d'exemple, la signification d'une phrase change totalement avec le changement de l'une de ses particules. Nous décrivons dans ce travail le processus de création et la méthodologie utilisés pour construire cette ressource. Cette dernière est constituée de 315 particules avec un ensemble de propriétés morpho-syntaxiques.

ABSTRACT. The technological development and the advanced growing of Arabic Natural Language Processing (ANLP) field offer advanced applications and lead to a particular interest in Language Resources. This paper focuses on Arabic particles resource development. This choice is explained by the importance of particles in sentences structure. As such, the meaning of a sentence changes completely with the change of one of its particles. Specifically, we describe the building process and methodology used for the resource that consists of 315 particles. Each one of them is detailed with its morpho-syntactic features.

MOTS-CLÉS : Langue arabe, lexique linguistique, particules Arabe, LMF.

KEYWORDS: Arabic language, linguistic lexicon, Arabic particles, LMF.

1. Introduction

In the high tech area (Smartphone, tablet, 4G, cloud computing, etc.) where the Information and communication technology have reached the maturity stage, Natural Language Processing (NLP) - at the intersection of three disciplines: linguistics, Information Technology and Statistics - is a growing field that offer advanced applications such as Information Retrieval, Machine Translation, Text Briefing and Question-Answering.

NLP applications, like any other development programs, must respect during the development process, the software quality assurance requirements[1] (design and coding requirements, code control, testing, change and release management, etc.) to ensure portability, interoperability and reusability of the application, and offer a product working in all platforms and compatible with any data format. In reality, many NLP applications suffer from the non-compliance with these exigencies (Nancy Ide, 2008).

These quality imperatives have promoted the importance given in recent years to both NLP applications and Language Resources (LRs) with the aim to ensure effective interoperability, because they are certainly a decisive element in this chain (Nancy Ide and James Pustejovsky, 2010).

This interest in LRs (mono or multi-lingual text, treebanks, dialogues, dictionaries, terminologies, ontologies) prompted experts to work in this specialized research domain and

[1] 730-2014 - IEEE Standard for Software Quality Assurance Processes.

to define them as data language available in an electronic format, and used for the development of language processing systems. LRs can be classified into two categories: corpora and lexicon (Bente Maegaard and al., 2006). As an example of corpora, we enumerate: The Corpus of Contemporary Arabic (Latifa Al-Sulaiti and Eric Atwell, 2006) and Quranic Arabic Corpus (Kais Dukes and Nizar Habash, 2010). As an example of lexicon, we enumerate: Buckwalter's list of Arabic roots (Tim Buckwalter, 2004) and Bilingual Dictionaries for Colloquial Arabic Dialects (David Graff and Mohamed Maamouri, 2012).

The recognition of the crucial role of LRs in the NLP sphere has led to the expansion and the revival of research and development in the field of language industries. This progress is promoted by the LRs community (Nicoletta Calzolari, 2008) formed by Distributors (LDC, ELRA, OLAC, NICT)[2], Research projects (CLARIN, FLaReNet, PAROLE, SIMPLE, WRITE, NEMLAR)[2], alliances and associations (META-NET, EAGLES, ELAN, NERC, RELATOR, AFNLP)[2] and conferences (LREC, ICGL, TELRI)[2].

In spite of this enlargement in the LRs community, many difficulties are encountered in the use of available LRs in a different environment than the original one. These troubles are due to the non-respect during the development process of some fundamental interoperability guides (Nancy Ide and James Pustejovsky, 2010):

- Availability: Accessibility (free vs. proprietary), affordability (resources at a very high cost) and customizability (the degree of manipulability of resources)
- Portability: ability to serve in different platforms
- Usability: application programming interfaces (APIs) implementation
- Standardization: Representation format that allows the coexistence of resources from different sources.

Arabic LRs as part of the Arabic Natural Language Processing (ANLP), present more challenges (Ali Farghaly and Kkaled Shaalan, 2009) than interoperability instructions due to Arabic language structure as a semitic, highly structured and derivational language with rich templatic morphology.

The most important challenge is shortage in Arabic LRs caused by two main factors. The first one is the independence between LRs and applications. For example, in the Buckwalter Arabic Morphological Analyzer, some resources are integrated in the code such as the integration of Arabic letters in `romanizeWord()` method, practice which makes the resource unusable outside the analyzer. The second one is the proprietary aspect of LRs such as those offered by distributors like Linguistic Data Consortium[3] (LDC) and European Language Resources Association[4] (ELRA). Additionally, even those freely available, they suffer from some constraints like the specificity of locally developed resources or the lack of details and linguistic information, except a few ones.

Thus, our main purpose in the current work is to contribute in the development of a set of Arabic LRs complying with the four mentioned interoperability rules (availability, portability, usability and standardization), with the aim of their wide distribution and use in ANLP research. We note that we haven't found similar free resource.

As such, the primary objective of this paper is to present particles resource with these methods, and a usage example for educational purposes. The particles choice as one of three categories of the Arabic word (noun, verb and particle اسم وفعل وحرف - Aism ,fiEl, Harf[5]) (Owens, Jonathan, 1989) is due to the non availability of particles resource meeting the above LRs requirement and as a completion of the previous development on the Arabic alphabet and

[2] Details in Abbreviation list in appendis A

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

[4] http://www.elra.info/

[5] According to Buckwalter Arabic transliteration.

affixes (Taoufik Loukili and Karim Bouzoubaa, 2011). It is natural to work on the particles before working later on the other two great important categories that are nouns and verbs.
In addition, the use of particles resource may be possible in different levels and by different tools in ANLP systems such as:

- Morphosyntactic analysis: Arabic sentences contain most of the time particles. Therefore, the knowledge of their morphosyntactic features makes this task easier;
- Corpus annotation: annotating particles in a corpus becomes simple with the availability of a particles lexicon giving annotations.

The particles data are designed by the collection of Arabic particles with their morphosyntactic features and their classification into three subsets (particle, special nouns and special verbs) according to the Arabic language word categories (محمد الجوهرى, 2013).
The structure of this paper is as follows: in the second part we define Arabic particles, in the third section we discuss Lexical Markup Framework (LMF: the ISO 24613 standard) as a data representation format used to hold Arabic particles, in the fourth section we examine particles resource with an example of use and we conclude with a summary and future works.

2. Arabic particles

Arabic words are divided into three categories: noun, verb and particle (عبد الوهاب حسن, 2011). Nouns are known by symptoms like:

- kasra acceptance (الخفض), for example : "In the neighbor's house - في منزلِ جاري - fy manzili jAry", the word "manzili - منزلِ" accept Kasra vowel
- double vowels acceptance (التنوين), for example : "This is a key - هذا مفتاحٌ - ha*A miftAHN", the word " miftAHN - مفتاحٌ " take dammatayn vowel
- definite article acceptance (دخول الألف واللام), for example : "the weather is clean - الجو صحو - Aljawu SaHwN", the word "Aljawu - الجو" accept the definite article.

Verbs are known by:

- qad acceptance (قد), for example : "May your hearing - قد سمعك - qad samiEaka", the word "samiEa - سمع" appear after "qad - قد "
- si~yn acceptance (السين), for example : "I will Wandering - سأتجول - sa>atajawalu", the word ">atajawalu - أتجول" is preceded by " si~yn - س"
- sawofa acceptance (سوف), for example : " you will succeed - تنجح سوف- sawofa tanjaHu ", the word "tanjaHu - تنجح " is preceded by "sawofa - سوف"
- ta' altanyth acceptance (تاء التأنيث), for example : "she wrote a wonderful article - كَتَبَتْ مَقالاً رائعاً - katabat maqaAlF raA}iEaAF ", the word "kataba - كَتَبَ" is terminated by "ta' altanyth - تاء التأنيث ".

Particles are words to which noun and verb symptoms cannot apply. They are divided into two categories: building particles (حروف المباني - Horowf AlmabAny) and meaning particles (حروف المعاني - Horowf AlmaEAny) (حفني ناصف and al., 2008). Building particles are the alphabet letters "ا، ب، ت، ث، ج، ح، خ" and meaning particles are what cannot be understood on its own and its shape does not change (مبنية) such as "من، في، بلى، حتى".
Meaning particles, that are the subject of our work in this article, are categorized in the Arabic literature according to several criteria:

- In terms of the work, Meaning particles are, worker (حروف عاملة) [Horowf EAmilap] such as"إن وأخواتها" or Not worker (حروف غير عاملة) [Horowf gayr EAmilap] such as "بلى، نعم".
- In terms of association, there is a distinction between particles which combine only with names (مختصة بالأسماء) [moxtaS~ap biAl>asmA'] like "preposition - حروف الجر", only with verbs (مختصة بالأفعال) [moxtaS~ap biAlOfEaAl] like "accusative particles -

حروف النصب" and which combine with both names and verbs [mo$tarakap lil>asmaA' wAl>afEaAl] (مشتركة للأسماء والأفعال) like "conjunction -حروف العطف ".

- Meaning particles are also divided into groups of several meanings depending on the moral purpose, to form classes with a common meaning:
    - Answer particle (حرف جواب) like : أجل، إي، بلى، نعم، لا[>ajal, <iy, balaY, naEam, laA]
    - Negative particle (حرف نفي) like : إنْ، لات، لا، ما، لنْ، لمّا، لم [Ino, laA, maA, lano, lam~aA, lam]
    - Conditional particle (حرف شرط) like : إنْ، إذ ما، لو، لولا، لوما[lawmaA, lawlaA, law, <i*maA, <ino]
    - Exhortation particle (حرف تحضيض) like: ألا، هلاّ، لولا، لوما [lawmaA, lawlaA, halA, >alaA]
    - Future particle (حرف استقبال) like: السّين، سوف، أنْ، إن، لنْ، هل [hal, lano, <in, >an, sawfa, Alsi~yn]
- Some authors divide particles into five sections: mono, bi, tri, quad, and quintet of characters.

In our design we chose to split the particles resource into three modules, similar to Arabic word categories, that are called particle, special nouns and special verbs. This categorization is explained by the fact that there are some names or verbs having the particles morphology. For example, the word "مَنْ" [mano] comply with particles symptoms and have particles morphology, but syntactically, it is a question name (اسم استفهام).

The multiple usage options of particles resource forces us to adopt a format that meets the above-mentioned interoperability guides (availability, portability, usability and standardization) through the adoption of a recognized standard such as the LMF that we present in the next section.

3. Lexical Markup Framework

The diversity and variety in norms and standards of LRs arise from the contribution of several organizations across research projects to propose standards, like BNC project [6], EAGLES/ISLE[7] (Nicoletta Calzolari and al., 2002) and LIRICS (The Linguistic Infrastructure for Interoperable Resources and Systems). This task of processing market standards is carried by ISO/TC 37/SC 4, the fourth subcommittee of the technical committees "Terminology and other language and content resources" in International Organization for Standardization, which develops ISO international standards for language resource management (LRM).

In this work we use the ISO 24613 (ISO 24613:2008, 2008) standard corresponding to the LMF norm (Lexical Markup Framework), in order to offer a LR in a highest degree of acceptability which ensures an easy exchange among NLP applications.

LMF is described by the International Organization for Standardization as a meta-model for representing data in lexical databases used with monolingual and multilingual computer applications, so as to provide mechanisms that allow the development and integration of a variety of electronic lexical resource types.

LMF support in its design other standards such as compatibility with the Unicode standard, and the use of linguistic information attribute-value pairs used in the ISO 12620 Data Category Registry (DCR)[8]. Additionally, LMF uses a subset of UML that is relevant for

[6] http://xml.coverpages.org/bnc-encoding2.html

[7] www.ilc.cnr.it/EAGLES96/isle/

[8] www.isocat.org

linguistic description complying with principles defined by the Object Management Group (OMG)[9].
LMF is composed of two blocks: LMF core package and LMF extensions. So, for particles resource representation, we use the morphological and syntactic LMF extensions.

4. Particles resource

In the respect of the LMF model, the particles resource is presented as follows:

Figure 1. *Particles XML data file extract*

This figure shows an excerpt of the XML file. According to the Arabic language structure, we have used Lemma and Word Form classes of LMF morphological extension and Syntactic Behaviour of LMF syntactic extension, which are used to describe, respectively, the written form of the lemma, the inflected forms properties and syntactic description. Detailed description of the syntactic behaviour is defined by the sub-categorization frame that can be exploited in the contextual exploration like sentences segmentation.
The resource contains three subcategories: particle, special nouns and special verbs.
In full respect of the Arabic language peculiarities, the “particle” sub-group contains the following common meaning Arabic particles: inceptive particle, coordinating conjunction, answer particle, preposition, future particle, conditional particle, amendment particle, exceptive particle, accusative particle, subordinating conjunction, vocative particle, negative particle, exhortation particle and supplemental particle. Figure 1 shows an example of the conjunction "About - عن" represented as a lexical entry.
Particles having name-like morphology such as relative pronoun, personal pronoun, demonstrative pronouns and interrogative pronouns are classified under the “special nouns” subcategory. Figure 1 shows an example of the personal pronoun "you - أنتم".
Moreover, particles having verb-like morphology such as certitude verbs, transposition verbs, hopefulness verbs and starting verbs, are classified under the “special verbs” subcategory.
The resource elaboration was done manually by linguists using the Arabic litterature such as "الجنى الداني في حروف المعاني" (المرادي، الحسن بن قاسم & فخر الدين قباوة و محمد نديم فاضل) to come up with a file containing 69 particles, 180 special nouns and 66 special verbs. The number 315 particles (which can reach ten thousand when particles are combined with affixes) may seem small, but it is large and most exhaustive compared to the number of Arabic particles. That file was structured and designed to respect the interoperability features mentioned above:

- portability: represented as an XML file, the particles resource can be used and exploited in any platform;

[9] www.omg.org

- standardization: respecting the LMF format, the particles resource is considered as a standard resource;
- usability: as explained below, the resource is usable not only as a file but also through a dedicated APIs

On another hand, to simplify the use of this resource without the need to fully understand its content and the structure of the LMF format, the Particles resources is also available through an API allowing ANLP developers accessing it by calling specific methods. This has been done in the context of the SAFAR platform (Seddik Sidrine and al., 2010) (Software Architecture For Arabic language pRocessing). Let us recall that SAFAR, is an open source platform dedicated for ANLP wich offers an integrated development environment dedicated to the development of Arabic NLP systems, offering tools such as morphological analyzers, morphological stemmers, sentence parsers, tokenization and transliteration. In addition, it offers methods to use and access resources such as corpora, alphabet, clitic, dictionary, ontology and particles.

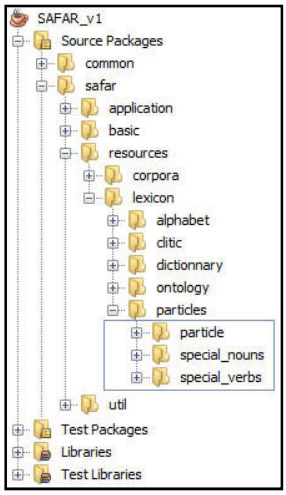

Figure 2. *Particles package in SAFAR plateforme*

The particles resource is then integrated as a data in the resources package (Figure 2). Under this package, we find the three categories of the particles resource (particle, special nouns and special verbs). Below is an extract of the most useful methods within this API.

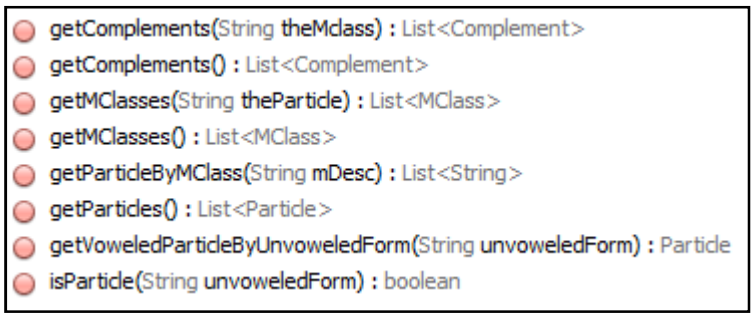

Figure 3. *Extract of Particles API* methods

For example, the following code illustrates the use of the method `isParticle()` that allows to test the existence of a word in particles data (if it's a particle).

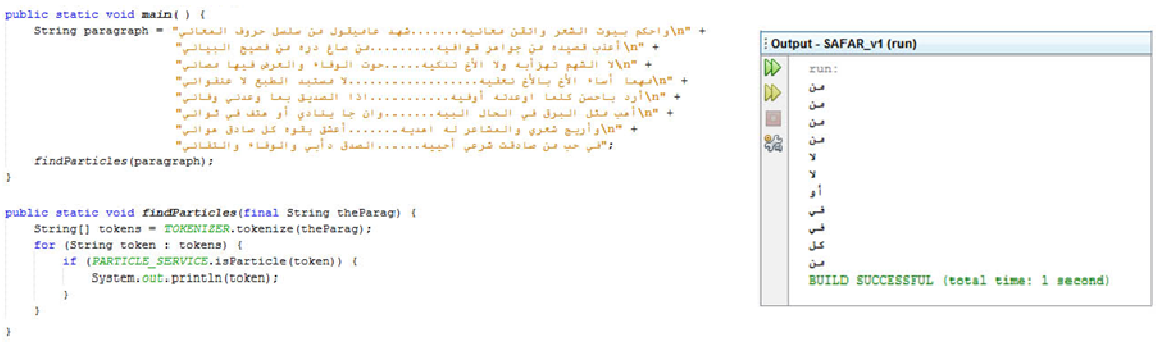

Figure 4. *`findParticles()` method*

Through a standard set of classes and methods, constituting a software library which serves as a channel which offer an indirect access to these resources, a programmer don't needs to know something else to make full use of the resources, without deepening in all internal details of these resources.
The second access method is the direct access to resources constituted by xml files exploiting in other tools.
The particles package can be easily used in several tools such as morphosyntactic analysis, automatic text generation, spell-checking and information retrieval applications. This example exposes the resource exploited for educational purposes.

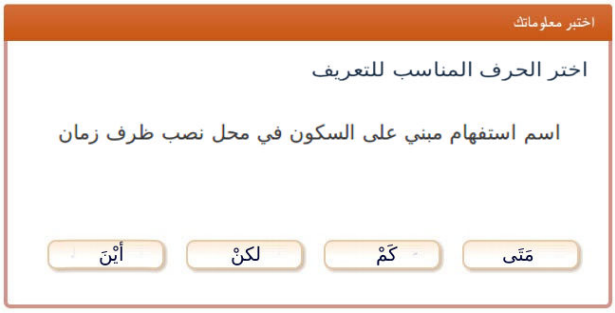

Figure 5. *Particles learning application*

An educational activity (Fanny Rinck and Thomas Lebarbé, 2005) is the implementation of a minimum objective. The educational objective of this activity is to assess the learner's knowledge about a central concept which is "particles" in an exercise. The activity consists for example to determine the most appropriate particle in a sentence or to fix the grammatical properties of some particles in a definition paragraph.
Learners in this exercise have to choose the correct answer from a list of choices.

5. Summary and future works

In this article, we describe the design process and production mechanisms of an Arabic particles resource with their morphosyntactic features, which consists of three sub-categories (particle, special nouns and special verbs) in the respect of interoperability guides in the context of the LMF-ISO 24613 standard.

In future works, we intend to develop other resources such as stop Lists, words roots and patterns lists in order to converge to dictionaries and lexicons so that to offer a wide range of usable resources in different platforms for different purposes.

References

Aïda Khemakhem, Imen Elleuch, Bilel Gargouri & Abdelhamid Ben Hamadou. 2009. Towards an automatic conversion approach of editorial Arabic dictionaries into LMF-ISO 24613 standardized model. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Ali Farghaly and Kkaled Shaalan. 2009. Arabic Natural Language Processing: Challenges and Solutions, the Association for Computing Machinery (ACM), Transactions on Asian Language Information Processing TALIP Vol 8, Issue 4, December.

Bente Maegaard, Steven Krauwer, Khalid Choukri & Lise Damsgaard Jorgensen. 2006. The BLARK concept and BLARK for Arabic. In Fifth International Conference on Language Resources and Evaluation, LREC'06.

David Graff and Mohamed Maamouri. 2012. Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects. Proceedings of the Second Language Resources and Evaluation Conference (LREC). p. 269-274, Istanbul, Turkey, May 21-27.

Fanny Rinck, Thomas Lebarbé. 2005. Constitution et exploitation pédagogique de ressources linguistiques pour un enseignement / apprentissage du discours rapport. Journées de la Linguistique de Corpus. Lorient, France.

Feten Baccar, Aïda Khemakhem, Bilel Gargouri, Kais Haddar, & Abdelhamid Ben Hamadou. 2008. Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe. In Proceedings of the 15eme Conference Annuelle sur le Traitement Automatique des Langues Naturelles (TALN'08).

ISO 24613:2008. 2008. Language resource management - Lexical markup framework (LMF). ISO/FDIS.

Kais Dukes and Nizar Habash. 2010, May. Morphological Annotation of Quranic Arabic. In LREC.

Latifa Al-Sulaiti and Eric Atwell. 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, no. 2, pp. 135-171, John Benjamins Publishing Company.

Nancy Ide. 2008. Global Interoperability: How Can We Get There?. The First International Conference on Global Interoperability for Language Resources, ICGL.

Nancy Ide and James Pustejovsky. 2010. What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. The Second International Conference on Global Interoperability for Language Resources, ICGL.

Nicoletta Calzolari and Ralph Grishman and Martha Stone Palmer. 2002. Standards & best practice for multilingual computational lexicons: ISLE MILE and more. Proceedings of the Language Resources and Evaluation Conference, Gran Canaria.

Nicoletta Calzolari. 2008. Approaches towards a "Lexical Web": the role of Interoperability. The First International Conference on Global Interoperability for Language Resources, ICGL

Owens Jonathan. 1989. The syntactic Basis of Arabic word classification. Arabica, 211-234.

Seddik Sidrine and Younes Souteh and Karim Bouzoubaa and Taoufik Loukili. 2010. SAFAR: vers une plateforme ouverte pour le traitement automatique de la langue Arabe. in the Special Issue on "Advances in Arabic Language Processing" for the International Journal on Information and Communication Technologies (IJICT), Serial Publications, June 2010, 11:2533-2541.

Taoufik Loukili and Karim Bouzoubaa. 2011. Structuration et Standardisation des ressources linguistiques de l'Arabe cas de l'alphabet, préfixes et suffixes. 3ème édition des Journées Doctorales en Technologies de l'Information et de la Communication, ENSA de Tanger.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.

المرادي، الحسن بن قاسم & ,فخر الدين قباوة و محمد نديم فاضل. الجنى الداني في حروف المعاني.

حفني ناصف ومحمد دياب ومصطفى طموم ومحمود عمر وسلطان محمد والشيخ محمد علي طه الدرة. 2008. قواعد اللغة العربية. مكتبة الآداب الطبعة الأولى.

عبد الوهاب حسن حمد احمد. 2011. وظيفة البناء الصرفي. جامعة بابل – العراق.

الجوهرى 2013.أنواع الجمل والقوالب النحوية |Al-Madinah Language Studies .مجلة جامعة المدينة العالمية لعلوم اللغة-مصر 25.25.

Appendix A: List of abbreviations

- **LDC**: Linguistic Data Consortium

- **ELRA**: European Language Resources Association
- **OLAC**: Open Language Archives Community
- **NICT**: National Institute of Information and Communication Technology
- **CLARIN**: Common Language Resources and Technologies Infrastructure
- **FLaReNet**: Fostering Language Resources Network
- **PAROLE**: Preparatory Action for Linguistic Resources Organisation for Language Engineering
- **SIMPLE**: Semantic Information for Multifunctional Plurilingual Lexica
- **WRITE**: Written Resources Infrastructure, Technology and Evaluation
- **NEMLAR**: Network for Euro-Mediterranean LAnguage Resources
- **META-NET**: Multilingual Europe Technology Alliance Network
- **EAGLES**: Expert Advisory Group on Language Engineering Standards
- **ELAN**: European Language Activity Network
- **NERC**: Network of European Reference Corpora
- **RELATOR**: European Network of Repositories for Linguistic Resources
- **AFNLP**: Asian Federation of Natural Language Processing
- **LREC**: Language Resources and Evaluation Conference
- **ICGL**: International Conference on Global Interoperability
- **TELRI**: Trans-European Language Resources Infrastructure