POS tagging in Amazighe using tokenization and n-gram character feature set

Mohamed Outahajala (1, 4), Yassine Benajiba (2), Paolo Rosso (3), Lahbib Zenkouar (4)
(1) Royal Institut for Amazighe Culture, Morocco,
(2) Philips Research North America, Briacliff Manor, USA,
(3) Natural Language Engineering Lab – EliRF, DSIC, Universidad Politécnica de Valencia, Spain,
(4) Ecole Mohammadia d'Ingénieurs, Morocco,

outahajala@ircam.ma, yassine.benajiba@philips.com, prosso@dsic.upv.es, zenkouar@emi.ac.ma

Résumé

L'objectif de cet article est de présenter le premier étiqueteur grammatical Amazighe. Très peu de ressources ont été développé pour l'Amazighe et nous croyons que le développement d'un outil d'étiquetage grammatical est la première étape dont on a besoin pour faire le traitement automatique de textes. Afin d'atteindre cet objectif, nous avons formé deux modèles de classification de séquences en utilisant Support Vector Machines (SVM) et Conditional Random Fields (CRFs) en utilisant une phase de segmentation. Nous avons utilisé la technique de 10 fois validation croisée pour évaluer notre approche. Les résultats montrent que les performances des SVMs et des CRFs sont très comparables. Dans l'ensemble, les SVMs ont légèrement dépassé les CRFs au niveau des échantillons (92,58% contre 92,14%) et la moyenne de précision des CRFs dépasse celle des SVMs (89,48% contre 89,29%). Ces résultats sont très prometteurs étant donné que nous avons utilisé un corpus de seulement ~ 20k jetons.

Abstract

The aim of this paper is to present the first Amazighe POS tagger. Very few linguistic resources have been developed so far for Amazighe and we believe that the development of a POS tagger tool is the first step needed for automatic text processing. In order to achieve this endeavor, we have trained two sequence classification models using Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) after using a tokenization step. We have used the 10-fold technique to evaluate our approach. Results show that the performance of SVMs and CRFs are very comparable. Across the board, SVMs outperformed CRFs on the fold level (92.58% vs. 92.14%) and CRFs outperformed SVMs on the 10 folds average level (89.48% vs. 89.29%). These results are very promising considering that we have used a corpus of only ~20k tokens.

Keywords:

Étiquetage grammatical automatique, langue Amazighe, apprentissage supérvisé.
Automatic POS tagging, Amazighe language, supervised learning.

1 Introduction

The part-of-speech (POS) tagging consists of annotating each word in a sentence with its lexical category, i.e. part-of-speech. It is the first layer above the lexical level and the lowest level of syntactic analysis. Hence, all the NLP tasks dealing with higher linguistic levels resort to the POS tags, namely: phrase chunking; word sense disambiguation; grammatical function assignment (Cutting et al., 1992) and named entity recognition (Benajiba et al. 2010a; Benajiba et al. 2010b). In conjunction with partial parsing, POS-tagging is used in more complex tasks (Manning and Schütze, 1999) e.g.: lexical acquisition, information extraction, finding good indexing terms in information retrieval and question answering.

In the literature, proof is abound that the most effective approaches to build an automatic POS-tagger are based on supervised learning machines (See Section 2), i.e. relying on a manually annotated corpus and often other resources, such as dictionaries and word segmentation tools, to pre-process the text and extract features. Similarly, in our approach we use sequence classification techniques based on two state-of-art machine learning approaches, namely: SVMs and CRFs, to build our automatic POS-tagger. We use a ~20k tokens manually annotated corpus (Outahajala et al., 2011) to train our models and a very cheap feature set consisting of lexical context and character n-grams to help boost the performance.

The rest of the paper is organized as follows: in Section 2 we present related work on POS tagging techniques in other languages. Then, in Section 3 we give an overview on the Amazighe language and the employed tag set in the Amazighe corpus. In Section 4 we present the two supervised approaches based on SVMs and CRFs that have been employed for POS tagging. Section 5 describes experiments and discusses results. Finally, in Section 6 we draw some conclusions and describe the work to be done in the future.

2 Related work on POS tagging

The very first POS taggers were mainly rule-based systems. Building such systems requires a huge manual effort in order to handcraft the rules and to encode the linguistic knowledge which governs the order of their application. For instance, in 1970 Green and Robin (Greene, Rubin, 1971) developed a system named TAGGIT containing about 3,000 rules and achieving an accuracy of 77%. Later on, machine learning based POS-tagging proved to be both less laborious and more effective than the rule based ones. In the literature, many machine learning methods have been successfully applied for POS tagging, namely:

- The Hidden Markov Models (HMMs) (Charniak, 1993) whose states are tags or tuples of tags. For a bigram tagger, for instance, the states of the HMM are tags, transition probabilities are probabilities of a tag given the previous tag and emission probabilities are probabilities of a word given a tag;

- The transformation-based error driven system (Brill, 1995) consisting in assigning the most frequent tag to each word by using an annotation reference. It proceeds afterwards by selecting the rule that yields the greatest error. This process is iterated as long as the annotation results are not close enough to the annotation reference;

- The decision trees (Schmid, 1999) based on a decision support tool that uses a model and its possible consequences constructed on the basis of an annotation reference;

- The maximum entropy model (Ratnaparkhi, 1996) permitting the combination of diverse forms of contextual information without imposing any assumptions on training data, where the goal is to maximize the entropy of a distribution subject to certain constraints contained in the annotation reference.

- Learning algorithms that acquire a language model from a training corpus: based on previously learned examples, taggers depending on this approach decide on the tag to attribute to the word (Kudo, Matsumoto, 2000; Lafferty et al. 2001).

Results produced by machine learning taggers obtain about 95%-98% of correctly tagged words. There are also, hybrid methods that use both knowledge based and statistical resources.

Though these methods have good performance, the accuracy for unknown words is much lower than that for known words, and this is a non-negligible problem where training data is limited. The tag set size may vary widely. For instance, in POS tagging Arabic (Diab et al., 2004) the authors used a tagset containing 22 tags and 75 tags in (Diab et al., 2007).

3 Amazighe

In this section we give a brief introduction to the Amazighe language and describe the adopted tag set in our experiments in Subsections 3.1. and 3.2.

3.1 The Amazighe language

The Amazighe language is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is used by tens of millions of people in North Africa mainly for oral communication, and has been introduced in mass media and in the educational system in collaboration with several ministries in Morocco. It belongs to the Hamito-Semitic/“Afro-Asiatic” languages, with rich templatic morphology (Chafiq, 1991). In linguistic terms, the language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors (the orthographic details are discussed in (Ameur et al., 2006).

In Morocco, for instance, one may distinguish three major dialects: Tarifit in the North, Tamazight in the center and Tashlhit in the southern parts of the country; it is a composite of dialects of which none have been considered the national standard.

Due to its complex morphology as well as the use of the different dialects in its standardization (Tashlhit, Tarifit and Tamazight being the three more used ones), the Amazighe language presents interesting challenges for NLP researchers which need to be taken into account. Some of these characteristics are:

1. It does not have capitalization in its script.

2. Its writing system, Tifinaghe, is written from left to right (Ameur et al. 2006; Zenkouar, 2008).

3. It is a complex morphology language.

4. Nouns, quality names (adjectives), verbs, pronouns, adverbs, prepositions, focalizers, interjections, conjunctions, pronouns, particles and determinants consist of a single word occurring between blank spaces or punctuation marks. However, if a preposition or a parental noun is followed by a pronoun, both the preposition/parental noun and the following pronoun make a single whitespace-delimited string. For example: ⵖ (ɣr) “to, at” + ⵉ (i) “me (personal pronoun)” results into ⵖⴰⵔⵉ/ⵖⵓⵔⵉ (ɣari/ɣuri) “to me, at me, with me”.

5. It is not an exception when it comes to POS ambiguity, i.e. same surface form might be tagged with a different POS tag depending on how it has been used in the sentence. For instance, ⵉⴳⵯⵔⴰ (ig°ra[1]) may have many meanings; as a verb, it means ‘lag behind’ while as a noun it refers to the plural noun of ⴰⴳⵔⵓ (agru) meaning ‘a frog’. Some stop words such as “ⴷ” (d) might function as a preposition, a coordination conjunction, a predicate particle or an orientation particle.

6. Like most of the languages which have only recently started being investigated for the NLP tasks, Amazighe lacks annotated corpora and tools and still suffers from the scarcity of language processing tools and resources.

3.2 Our tag set

Defining the adequate tag set is a core task in building an automatic POS tagger. It aims at defining a processable tag set which is neither so large that it can hurt the performance of the learning machines, nor so small that there is not enough information to be used by the potential federate systems. In (Outahajala et al., 2010), a tag set containing 13 elements (verb, noun, adverb…etc.) was developed. For each element we define morpho-syntactic features and two common attributes: “wd” for “word” and “lem” for “lemma”, whose values depend on the lexical item in question. The defined Amazighe elements and their attributes are set out in (Outahajala et al., 2011). The utilized tag set comprises 13 tags representing the major parts of speech in Amazighe, as it is summarized in Table 1. This tag set is derived from the larger one presented in (Outahajala et al., 2010). Gender, person and number information have not

[1] The amazighe Latin transliteration used in this paper is the one defined in (Outahajala et al., 2010)

been included in the tag set and were considered as a separate investigation subject to be pursued in the future.

Labeled class	Designation
V	Verb
N	Noun
A	Quality name/Adjective
AD	Adverb
C	Conjunction
D	Determinant
S	Preposition
FOC	Focalizer mechanism
I	Interjection
P	Pronoun
PR	Particle
R	Residual (foreign, number, date, currency, mathematical and other)
F	Punctuation

Table 1. tag set.

4 Supervised learning for POS tagging

In this section we describe the theoretical foundations of supervised learning in general and of SVMs and CRFs in particular, being proved to give good results for sequence classification (Kudu, Matsomoto, 2000; Lafferty et al., 2001).

In this paper, stems + affixes and punctuation marks are referred to as tokens.

4.1 Supervised learning

In supervised learning the goal is to learn a function:

$$h : X \rightarrow Y \qquad (1)$$

Where $x \in X$ are inputs and $y \in Y$ are outputs. The input objects are called instances, or examples, and they can be any kind of object, depending on the particular learning task: in NLP they could be for example documents to classify, strings of words to tag with POS-sequences which is our case. Depending on the nature of the output space Y, learning tasks can be categorized into several types:

- Binary classification: $Y = \{-1,+1\}$;
- Multiclass classification: $Y = \{1, \ldots ,K\}$ (finite set of labels);

- Regression: Y = R;
- Structured prediction: here the outputs in Y are complex. For example, in a sequence labeling task such as POS-tagging, Y = {1…,K}, i.e. the output is a sequence of labels of length n equal to the length of the input string.

4.2 Support Vector Machines

SVMs were first introduced by Vapnik (Vapnik, 1995); they are known for their good generalization performance and have been used for different recognition problems. For instance, in NLP SVMs are applied to text categorization (Kudu, Matsomoto, 2000), name entity recognition (Benajiba et al., 2010), base phrase chunking (Diab et al., 2007) and others. Many POS taggers based on SVMs have been achieved for many languages, such as: Arabic (Diab et al., 2004; Diab et al., 2007), Bengali (Ekbal, Bandyopadhyay, 2008) etc .They are reported to have achieved a high accuracy without over fitting even with a large number of features. SVMs are also known for coping well with sparse and noisy data.

With respect to the task of POS tagging in Amazighe, the training process has been carried out by YamCha[2], an SVM based toolkit. For classification, we have used the TinySVM-0.09[3] classifier, a publicly available toolkit for the problem of pattern recognition.

4.3 Conditional Random Fields

CRFs are undirected graph models. They are a generalization of Maximum Entropy Markov Models (MEMMs) and are oriented toward segmenting and labeling data (Lafferty et al., 2001). Conditional model specifies the probabilities of possible label sequences given an observation sequence. In addition to having the advantages of MEMMs, CRFs also overcome the label bias problem. We can think of CRFs as a finite state model with unnormalized transition probabilities. CRFs are applied to many NLP fields name entity recognition (Benajiba et al., 2010), shallow parsing (Sha, Pereira, 2003), information extraction from tables (Pinto et al., 2003). Dealing with POS tagging CRFs were used for many languages, such as Amharic (Adafre, 2005) and Tamil (Lakshmana, Geetha, 2009).

We have used CRF++[4], an open source implementation of Conditional Random Fields for segmenting and labeling data, using the same data set as the one used with YamCha.

[2] http://chasen.org/~taku/software/yamcha/

[3] http://chasen.org/~taku/software/TinySVM/

[4] http://crfpp.sourceforge.net/

5 Experiments and Error analysis

5.1 Corpus

Our corpus consists of a list of texts extracted from a variety of sources such as: the Amazighe version of IRCAM's web site[5], the periodical "Inghmisn n usinag" (IRCAM newsletter) and primary school textbooks. The texts are annotated using AnCoraPipe tool (Bertran et al., 2008). Annotation speed of this corpus was between 80 and 120 tokens/hour. Randomly chosen texts were revised by different annotators. On the basis of the revised texts, inter-annotator agreement is 94.98%. Common remarks and corrections were generalized to the whole corpora in the second validation by a different annotator.

The input format for YamCha and CRF++ is the same (see Figure 1 and Figure 2), where the first column is for tokens, the last column is for the labeled class and if any features are used (see Figure 2) they are listed in the columns in between.

Here is an example, where we don't use segmentation of the input format for the sentence "ar as ttHyyaln i tmGra ann sg usgg°as lli izrin." [English translation: "They were preparing for the weddings since the last year"]:

ar	PR
as	S_P
ttHyyaln	V
I	S
tmGra	N
ann	D
sg	S
usgg°as	N
lli	P
izrin	V
.	F

Fig.1. An extract from the training corpus

ar	a	-	-	-	r	-	-	-	PR
as	a	-	-	-	s	-	-	-	S_P
ttHyyaln	t	tt	ttH	ttHy	n	ln	aln	yaln	V
i	-	-	-	-	-	-	-	-	S
tmGra	t	tm	tmG	tmGr	a	ra	Gra	mGra	N
ann	a	an	-	-	n	nn	-	-	D
sg	s	-	-	-	g	-	-	-	S
usgg°as	u	us	usg	usgg	s	as	°as	g°as	N
lli	l	ll	-	-	i	li	-	-	P
izrin	i	iz	izr	izri	n	in	rin	zrin	V
.	-	-	-	-	-	-	-	-	F

Fig.2. An extract from the training corpus using lexical features

[5] http://www.ircam.ma/

In this paper, we explore two experiments sets. Both of them are based on SVMs and CRFs with and without lexical features (see subsection 5.2). In the first experiment set, we do not segment the compound words. In this set we use “S_P” and “N_P” to respectively designate prepositions and kingship nouns when both are followed by personal pronouns. In the second experiment set, we segment prepositions and kingship nouns when followed by pronouns. However, this is a problematic case since reverse function is not deterministic. We will use the most frequent words in the corpus. For instance, the union of either the two morphemes “dg” and “s” can give either “digs” or “dags”, meaning [in it]. Hence, once we split the compound word, e.g. “digs”, into its component morphemes, i.e. “dg” and “s”, and given that it is not possible to compute the original form after POS tagging, we will use “digs” since it is the most used form in the corpus. In all our experiments, we have used two tag sets: the one presented above (see Table 1) and this same tag set plus the two tags “S_P” and “N_P” corresponding to prepositions and kingship nouns respectively when followed by pronouns.

5.2 Features

In this paper, we explore two features sets. Both of them are based on the actual text and are very cheap to extract. In the first feature set (illustrated in Figure 1), we use: the surrounding words and their POS-tags in a window of -/+2. In the second feature set (illustrated in Figure 2), we add to the first feature set character n-gram feature which consists of the last and first i character n-gram, with i spanning from 1 to 4.

5.3 10-fold experiments

In our first experiments about POS tagging (Outahajala et al. 2011b), we have shown that learning curve is increasing with training corpus size. In this experiment set, we have run 10-fold cross validation over the corpus, i.e., training on 90% of sentences and tagging the remaining 10%, with the experiment repeated 10 times, each time taking a different slice of the corpus.

Fold#	SVMs	SVMs (with lexical features)	CRFs	CRFs (with lexical features)
0	81,01	86,86	83,19	86,95
1	76,02	83,86	80,7	84,98
2	85,64	**91,66**	87	90,86
3	82,56	88,34	86,45	88,58
4	83,55	88,24	85,8	88,87
5	83,28	89,99	86,24	90,48
6	76,59	85,38	79,98	85,38
7	79,07	86,6	81,79	87,96
8	87,35	91,38	88,88	91,14
9	84,64	90,41	86,79	**91,35**
AVG	**81,97**	**88,27**	**84,68**	**88,66**

Table 2. 10-fold cross validation results.

By splitting the compound words, namely kingship nouns and prepositions when followed by pronouns, we obtained better results as shown in Table 4:

Fold#	SVMs	SVMs (with lexical features)	CRFs	CRFs (with lexical features)
0	82,85	87,94	84,46	87,31
1	78,27	85,06	81,55	85,9
2	87,59	**92,58**	87,9	91,42
3	83,95	89,62	87,39	89,22
4	85,06	89,02	86,93	89,26
5	86,08	91,38	87,6	91,62
6	79,27	86,42	82,9	87,18
7	81,34	86,96	83,69	88,96
8	**88,54**	92,47	**89,32**	91,79
9	86,45	91,49	88,65	**92,14**
AVG	**83,93**	**89,29**	**86,01**	**89,48**

Table 3. 10-fold cross validation results using tokenization.

5.4 Experiments and Result Discussion

For a better understanding of the behavior of our system, we have examined the confusion matrix for the experiment which gave the highest accuracy. The analysis of the confusion matrix presents all the misclassified tags as shown in Tables 4 and 5.

Analyzing the most frequent errors in the two confusion matrices given by SVMs and CRFs, we found that adjectives are frequently tagged as nouns. This is due to the fact that adjectives may act as nouns. In line with this, many Amazighe linguists gave the name of quality nouns to adjectives. However, by dropping the distinction between nouns and adjectives we obtained an improvement of 0.73 and a better score of 90.02% using 10 fold cross validation. However, by doing the same experiment with CRFs, we obtained an improvement of 0.77 and a better score of 90.25%.

Error rate of pronouns is also high due to the large overlap between them and determinants. Another common source of errors is verbs. The POS tagger based on CRFs tagged 4.1% of verbs as nouns and adjectives and 1.6% as prepositions, whereas the POS tagger based on SVMs tagged 5.7% of verbs as nouns and adjectives. Besides SVMs based POS tagger has better results in tagging pronouns, determinants, adverbs, focalizers and particles.

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	93,1	0,3	1,8	0,6	3,9	0	0	0	0	0	0,3	0	0
A	18,2	63,6	18,2	0	0	0	0	0	0	0	0	0	0
V	5,4	0,3	93	0	0	0,7	0	0	0,7	0	0	0	0
P	0,7	0	0,7	91	5,5	0,7	0,7	0	0,7	0	0	0	0
D	3,3	0	1,1	9,9	84,6	0	0	1,1	0	0	0	0	0
S	0,5	0	1	0,5	0	94	2,1	0,5	1,6	0	0	0	0
C	0	0	0	2,1	2,1	2,1	83,3	4,2	4,2	2,1	0	0	0
AD	23,2	0	7,1	1,8	1,8	3,6	1,8	60,7	0	0	0	0	0
PR	0	0	0	0	1,9	0,6	0	0,6	96,8	0	0	0	0
FOC	0	0	0	0	40	0	0	0	0	60	0	0	0
F	0	0	0	0,2	0	0	0	0	0	0	99,8	0	0
I	36,4	0	0	0	0	0	0	0	18,2	0	0	45,4	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4. Confusion matrix in percentage when using SVMs with lexical features.

	N	A	V	P	D	S	C	AD	PR	FOC	F	I	R
N	94,6	2,4	2,1	0,2	0,3	0,1	0,2	0	0	0	0	0	0,1
A	12,6	82,3	4,6	0	0	0	0	0	0	0,6	0	0	0
V	2,6	1,5	93,3	0	0,4	1,5	0	0,4	0,4	0	0	0	0
P	3,7	0	0	75	13,9	0,9	0,9	0,9	3,7	0,9	0	0	0
D	2,4	0	0	4,8	82,5	0	0	2,4	7,9	0	0	0	0
S	0	0	0,2	0,3	0	99	0,5	0	0	0	0	0	0
C	1,7	0	0,6	0	0,6	2,9	91,4	0	2,9	0	0	0	0
AD	23,8	0	0	9,5	0	9,5	14,3	42,9	0	0	0	0	0
PR	0	0	1,1	1,1	2,3	1,1	9,2	1,1	83,9	0	0	0	0
FOC	0	0	0	0	0	0	0	0	50	50	0	0	0
F	0	0	0	0	0	0	0	0	0	0	100	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0
R	3,1	0	1,6	1,6	0	4,7	0	0	4,7	0	6,3	0	78,1

Table 5. Confusion matrix in percentage when using CRFs with lexical features.

Some particles are confused for example conjunctions like “d” which has many eventual tags depending on the context. For instance, in the sentences bellow, the word “d” might be:

- A coordination conjunction: “tamaziGt d tiknulujiyin timaynutin” [Amazighe and information technologies];
- A preposition: “iman d ubrid” [he went with the road];
- A predication particle: “d argaz” [he is a man];
- Or an orientation particle: “asi d tikint tamjahdit” [bring a large bowl].

Analyzing training and test sets, we observed that unknown words in the test set are important due to small size of the data set; also some errors still exist in the hand annotated corpora. Overall, misclassified words are unseen in the training set.

6 Conclusions and Further Work

In this paper we have tried to describe the morpho-syntactic features of the Amazighe language. We have addressed the design of two tag sets and two POS taggers based on SVMs and CRFs. The obtained accuracy achieved is 92.58%. We have used the 10-fold technique to further validate our results. In this way, the POS tagger based on CRFs achieves 89.48% whereas the POS tagger based on SVMs while using lexical features yields the accuracy of 89.29% based on a small manually annotated corpus of ~20k tokens.

We are currently trying to improve the performance of the POS tagger by using additional features and more annotated data based on semi-supervised techniques and active learning. In addition, we are planning to approach base phrase chunking by hand labeling the already annotated corpus with morphology information.

Acknowledgements

We would like to thank all IRCAM researchers for their valuable assistance. The work of the third author was funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

References

Adafre, S. F. (2005), Part of Speech tagging for Amharic using Conditional Random Fields. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pp. 47-54.

Ameur, M., Bouhjar, A., Boukhris, F. Boukouss, A., Boumalk, A., Elmedlaoui, M., Iazzi. (2006), Graphie et orthographe de l’Amazighee. Publications de l’IRCAM.

Benajiba Y., Diab M., Rosso P. (2010a), Arabic Named Entity Recognition: A Feature-Driven Study. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934. DOI: 10.1109/TASL.2009.2019927.

Benajiba Y., Zitouni I., Diab M., Rosso P. (2010b), Arabic Named Entity Recognition: Using Features Extracted from Noisy Data. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Uppsala, Sweden, July 11-16, pp. 281-285.

Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008), AnCoraPipe A tool for multilevel annotation. Procesamiento del lenguaje Natural, nº 41. Madrid, Spain.

Brants, T. (2000), TnT - A Statistical Part-of-Speech Tagger.

Brill, E. (1995), Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging

Chafiq, M. (1991), أربعة وأربعون درسا في الأمازيغية (Forty four lessons in Amazighe). éd. Arabo-africaines.

Charniak, E. (1993), Statistical Language Learning MIT Press, Cambridge

Cutting, D., Kupiec, J., Jan Pedersen, J. Sibun, P. (1992), Practical Part-of-Speech Tagger. Xerox Palo Alto Research Center.

Diab, M., Hacioglu, K., Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL)

Diab, M., Hacioglu, K., Jurafsky, D. (2007), Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter 9. Springer.

Ekbal, A.; Bandyopadhyay, S. (2008), Part of Speech Tagging in Bengali Using Support Vector Machine. In Information Technology, ICIT '08, pp. 106-111.

Greene, B.B., and Rubin, G.M. (1971), Automatic Grammatical Tagging of English. Department of Linguistics, Brown University, Providence, R.I.

Kudo, T., Matsumoto, Y. (2000), Use of Support Vector Learning for Chunk Identification.

Lafferty, J. McCallum, A. Pereira, F. (2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In proceedings of ICML-01, pp. 282-289.

Lakshmana Pandian S., Geetha T. V. (2009), CRF Models for Tamil Part of Speech Tagging and Chunking. In Proceeding ICCPOL '09. Springer-Verlag Berlin, Heidelberg

Manning, C., Schütze, H. (1999), Foundations of Statistical Natural Language Processing. The MIT Press.

Outahajala M., Zenkouar L., Rosso P., Martí A. (2010), Tagging Amazighe with AncoraPipe. In: Proc. Workshop on LR & HLT for Semitic Languages, 7th International Conference on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 52-56.

Outahajala M., Zenkouar L., Rosso P. (2011a), Building an annotated corpus for Amazighe. Will appear in Proceedings of 4th International Conference on Amazigh and ICT. Rabat, Morocco.

Outahajala M., Benajiba Y., Rosso P., Zenkouar L. (2011b), POS tagging in Amazighe using Support Vector Machines and Conditional Random Fields. Will appear in Proceedings of 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011.

Pinto, D., McCallum, A., Wei, X., Croft. W. B. (2003), Table extraction using conditional random fields. In SIGIR '03: Proceedings of the 26th annual international, pp. 235-242, New York, USA

Ratnaparkhi, A. (1996), A Maximum Entropy Model for Part-Of-Speech Tagging. In proceedings of EMNLP, Philadelphia, USA.

Schmid, H. (1999), Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Academic Publishers, Dordrecht, 13-26.

Sha, F. and Pereira F. (2003), Shallow Parsing with Conditional Random Fields. In Proc. of Human Language Technology.

Vapnik, Valdimir N. (1995), The Nature of Statistical Learning Theory. Springer Verlag, New York, USA.

Zenkouar L. (2008), Normes des technologies de l’information pour l’ancrage de l’écriture Amazighe, revue Etudes et Documents Berbères n°27, pp. 159-172.