N-class language model for Tunisian dialect automatic speech recognition system

Abir Masmoudi, Rim Laatar, Mariem Ellouze and Lamia Hadrich Belguith

University of Sfax, MIRACL Laboratory, Tunisia
masmoudiabir@gmail.com, laatar.rim@gmail.com,
Mariem.Ellouze@planet.tn, lamia.belguith@gmail.com

Abstract. Among the essential components of a speech recognition system comes the language model. This model purposes to define a probability distribution on sets of word sequences. In this context, we are interested in the statistical modeling of language, especially spontaneous speech recognition systems in the Tunisian dialect. Since this dialect suffered from the lack of data and resources, we propose to build an n-class language model that is based mainly on the integration of purely semantic data. In order to evaluate the model generated by our statistical language modeling system, we first calculated its perplexity, and in a second time we compared it to another model "n-gram" by using the SRILM tool. The result of the comparison between the two models, n-class model and n-gram model, proves that the predictive power of n-class model is better than that of n-gram model which presents a high value of its perplexity.

Keywords: language model, n-class, Tunisian dialect, speech recognition system.

1 Introduction

The speech recognition field remains a topic of current research. Several research efforts have been carried on in recent years to propose solutions in order to allow the automatic passage of a speech signal to the text. Today, speech recognition is integrated into concrete applications, widely known as human-machine oral dialogue applications. Among the major components of an automatic speech recognition system is the language model. This model purposes to define a probability distribution on sets of word sequences. In addition to the case of automatic speech recognition, the use of statistical language modeling intervenes for instance, in automatic translation, information search, text categorization or optical character recognition.

In the literature, several statistical approaches for language modeling are recognized as being the best performers in automatic recognition speech. These approaches are based on the estimation of probabilities n-grams or sequences of n-words. Among these approaches, we can state the n-grams model, n-class language model, factorial language model etc. Due to its simplicity and efficiency, the n-grams model constitutes the language model, which is most commonly used in the speech field. It is based on the assumption that the probability of the appearance of a word depends only on the history of some n-1 words that precede it. In practice, the estimation of this probability is very difficult. In fact, no learning corpus can make it possible to observe all the sequences of possible words. As a result, the basic idea of the n-grams models consists in considering only the sequences of words of length n**.** i.e. the calculation is approached by a limited history consisting of the n-1 preceding words. So, the major disadvantage of this modeling type leads to assigning a zero probability to any n-gram that has never been encountered in the learning corpus. In order to meet this requirement, other methods have emerged. Due to the lack of learning data, it is necessary to find a method that maximizes the amount of information. This corresponds to the appearance of n-class language model. The main idea of this model is to classify vocabulary words into lexical classes and to calculate the probability of a sequence of words such as the probability of a sequence of lexical classes [5].

The use of n-class language model is therefore justified by the following reasons: (1) the amount of learning data is reduced; (2) several words present similar behaviors. In this context, the use of n-class model is beneficial on several levels. On the one hand, the advantage of this method can be noticed in the fact that a word of a given class, not necessarily found in the learning corpus, inherits the probability of all the other representatives of its class. On the other hand, it is possible to add words to classes without the need to re-estimate the probabilities of the model. Thus, for n-class language model, classification methods can be based on syntactic information (common name, verb, preposition, etc.), semantics and also automatic classification methods [4].

Our present work aims to propose a method for the construction of a language model as part of the realization of Tunisian dialect speech recognition system for the Tunisian Railway Transport Network. However, the development of a statistical language model for spontaneous speech in the Tunisian dialect faces several difficulties. The main problem is the scarcity and even the lack of data in this dialect. Indeed, the availability of spontaneous data goes through a collection phase and a transcription of dialogue sessions in authentic conditions. This phase is extremely cumbersome and expensive to be implemented. Since our field of work is limited, there are several words with similar behavior (semantic or grammatical for example) but they do not have the same probability of appearance; their grouping in class will therefore be possible. For these reasons, we propose to build an n-class language model that is based mainly on the integration of purely semantic data. Indeed**,** our method will be used to create a language model based on semantic information for the creation of word classes.

This paper is organized as follows. Section 2 discusses the related work in the language model field and summarizes the main aspects of every work. Section 3 exposes the dataset used in our experiments, introduces our proposed model and presents our experiments and results. We finally draw some conclusions in Section 4.

2 Related Work

Several works are developed in the literature to classify vocabulary words for the construction of an n-type class language models. Below, an overview of the different existing works.

In the context of training classes of words, [9] proposes a simple word classification algorithm for statistical language modeling in speech recognition. The classification criterion used in this approach is the similarity of words. Indeed, the principle is based on the criterion of substitution or replacement. According to this algorithm, two words are similar since they can be substituted in the learning corpus [9]. According to this automatic word classification approach**,** the word accuracy rate was increased by 8.6% with a reduction in perplexity of about 6.9% [9].

The decision trees used in language modeling attempt to predict the next word from relevant questions that are all ways of extracting information from the history of the word [1]. The construction algorithm consists in successively selecting the questions that best suit the representation of the learning data. The general construction criterion is the minimization of the average entropy of the leaf distributions and therefore the minimization of uncertainty in the decisions made. This potentially very rich formalism can be applied wisely to quite varied fields. However, it remains rather expensive in calculations and requires expertise on several levels [8].

Brown's algorithm is commonly used in language modeling. Thus, in the application context of class-based language models, as proposed by [5], the modalities of the variables X and Y are identical and consist of vocabulary words. The criterion of grouping classes is based on the evaluation of the distance between each pair of classes. Thus, the distance between two classes C1 and C2 is none other than the reduction of mutual information generated by the eventual grouping of these two classes. The algorithm is thus quite expensive insofar as it requires a complete update of the distance matrix at each iteration.

The method proposed by [8] is essentially based on the principle of combining different sources of information at the class formation level. In his work, [8] uses two types of information: contextual information and prior information. The former is the most commonly used, corresponds to n-gram dependencies. This information can be collected not only at the words level, but also at the level of previously constructed classes of words [8]. It is fundamental to take into account the contextual information in order to better distribute the words into the classes. Thus, the use of contextual information is of interest in the context of improving the predictability of the model. It makes it possible to offer a better distribution of words into classes and thus, a more balanced distribution of distributions [8]. The second type, either semantic or syntactic information, is formalized by categories or grammars. In the approach proposed by [8], the used information a priori is extracted from a learning corpus labeled in grammatical categories.

The approach proposed by [24] is based on contextual information (left context and right context), so words that appear frequently in similar contexts should be assigned to the same class. According to [24], different vocabulary words are classified using the k-means algorithm. The particularity of this approach is based on the fact that the number of words in a class is set to k and if there is a class whose number of words is less than k then that class will be merged with another. The main advantage of this algorithm is its simplicity to find centroids and suddenly, the cost of merging words or classes becomes less expensive.

The approach developed by [2] proposes to integrate semantic information for the formation of word classes in the statistical language model of an automatic speech recognition system. This approach is based on a pivot language (called IF for Interchange Format), which represents the meaning of the sentence regardless of the language [2]. Thus, the criterion of choice of classes is guided by the definition of the pivot language and the most used concepts in the IF.

3 Method overview

Our proposed method consists of three fundamental phases, namely the construction and standardization of the corpus, the construction of the language model, and the evaluation of this model by calculating its perplexity rate. The first phase consists of three steps, such as recording, manual transcription and standardization. The second phase is made up of three steps, namely semantic labeling, word classification and language model calculation using the SRILM tool. The third phase deals with the evaluation of the constructed model by calculating its perplexity. These phases and steps are shown in Figure 1.

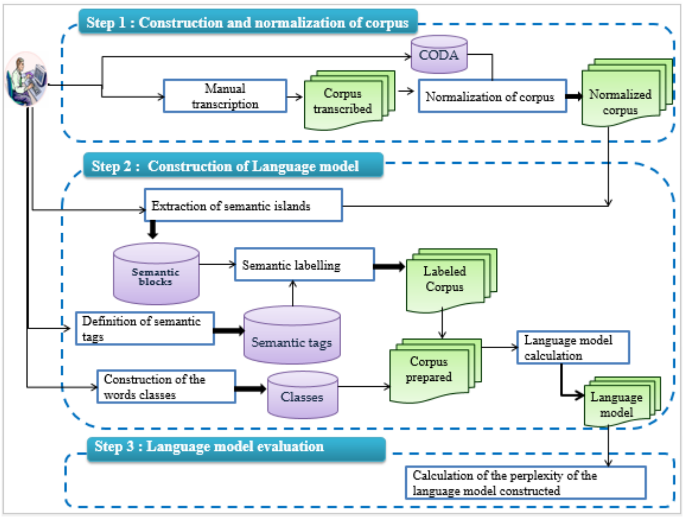

Step 1 : Construction and normalization of corpus
CODA
Manual transcription
Corpus transcribed
Normalization of corpus
Normalized corpus
Step 2 : Construction of Language model
Extraction of semantic islands
Semantic blocks
Semantic labelling
Labeled Corpus
Definition of semantic tags
Semantic tags
Construction of the words classes
Classes
Corpus prepared
Language model calculation
Language model
Step 3 : Language model evaluation
Calculation of the perplexity of the language model constructed

Fig. 1. Process of constructing a language model

3.1 Construction of TARIC: Tunisian Arabic Railway Interaction Corpus

We create our own corpus of real spoken dialogues corresponding to the information request task in railway stations in collaboration with the Tunisian National Railway Company (SNCFT)1. This corpus is called TARIC, for Tunisian Arabic Railway Interaction Corpus [16]. The main task of the TARIC corpus is information request in the Tunisian dialect about the railway services in a railway station. These requests are about consultation, train type, train schedule, train destination, train path, ticket price and ticket booking. The creation of the corpus was done based on three steps. The first one is the production of audio recordings; the second is the transcrip-

[1] http://www.sncft.com.tn/

tion of these recordings; and the third is the normalization of these transcriptions. In the following three sub-sections we will explain the process of creation of TARIC.

- **The Recordings:** The first step consisted in making audio recordings. We did that in the ticket offices of the Tunis railway station. We recorded conversations in which there was a request for information about such things as the train schedules, fares, bookings, etc. We obtained 20 hours of audio recordings.
- **The Transcription:** Once our recordings were ready, we transcribed them manually due to the absence of automatic tools for transcription for the Tunisian Dialect. This transcription was done by three university students. Our corpus consisted of several requests**,** which could be combined together during a dialogue between the staff and the client about railway services in the train station.
- **The normalization:** Unlike other languages, the Tunisian dialect has no written standard and systematic descriptions of its phonological, morphological, syntactic, semantic and lexical systems. Therefore, we developed our own orthographic guidelines to transcribe the spoken Tunisian dialect following previous work by [12] on developing a conventional orthography for dialectal Arabic – or CODA [25]. During the transcription of our corpora, we used the writing standards set in our normalization convention CODA in order to obtain coherent and consistent data.

This corpus consists of 1824 dialogues representing 6563 client statements and 5651 agent statements. Table 1 describes the characteristics of our corpus.

Table 1. Description of TARIC corpus used in our experiments

# of dialogues	1824
# of statements	21102
# of words	66082
# of customer statements	5651

3.2 Language model construction

The language model construction phase goes through four steps**,** namely building semantic blocks, semantic labeling, the formation of word classes and the calculation of language model using the tool SRILM. In what follows we will give some detail about these different steps.

- **Construction of semantic blocks:** The construction of semantic blocks consists of grouping one or more words into a single word that we call "semantic blocks". In our work, a semantic block is defined as a group of two or more words. Indeed**,** this pretreatment consists in adding a (-) between two or more words to build a single word. Among the words that can be grouped together to form a semantic block, we find "ماضي" followed by another word to indicate the time for example "ماضي ساعة" [1 PM]. Cities whose names are composed such as "برج سدرية " [name of Tunisian city]. "ما" [negation] followed by a verb with a negative form to express negation.

This step is necessary because it will be used for semantic labeling and later for the formation of word classes. Indeed, the main objective of this step is to give a better semantic value to words that may be insignificant and subsequently useless for our work.

- **Semantic labeling:** In order to obtain a labeled corpus, the semantic labeling step consists in giving a label for each single word or for each semantic block. Table 2 shows examples of words with the proper labels.

Table2. Examples of possible labels

Words/ semantic block	Labels
التْرَانْ, التْرِينُو	Concept-Train
لُورَارْ, تَوْقِيتْ,لَايْ-زُورَارْ	Concept-Hour
أَلَايْ-رُتُورْ	Ticket-Type
أَرْبَعَةْ, سَانْكُنْتْ, سِتْةْ, تْنَيْنْ	Nombre

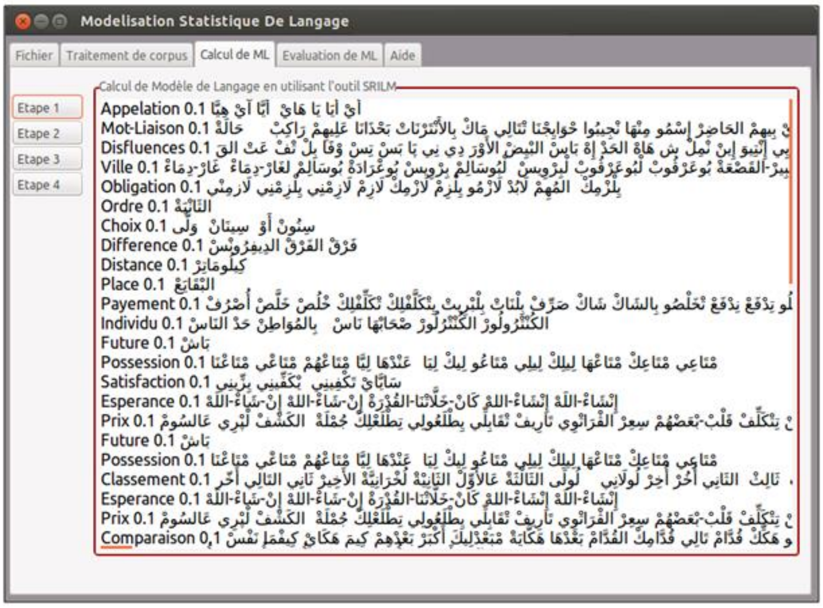
Thus, semantic labeling is not done word by word because we can find words that can have several meanings depending on the context in which they are used. Subsequently, the labeling of a word or a semantic block is done while taking into account his left and right neighbors in the statement.

- **Construction of semantic classes:** The present work being mainly dedicated to building a class-based language model, focuses essentially on the formation of semantic classes. In fact, a semantic class may correspond to a label or group of labels, whereas a label cannot belong to only one class. Table 3 presents some semantic classes.

Table 3. Examples of Semantic Classes

Semantic classes	Variants associated semantic tags
City	Destination-Station, Station…
Action	Concept-Departure, Concept-Arrivee
Response	Confirmation, Negation

After obtaining the list of semantic classes, as shown in Table 3, we can then directly associate each word of our corpus with the class to which it belongs. Figure 2 presents an extract of prepared corpus, of which each class contains the words (or semantic blocks) of which they belong.


Modelisation Statistique De Langage
Fichier
Traitement de corpus
Calcul de ML
Evaluation de ML
Aide
Calcul de Modèle de Langage en utilisant l'outil SRILM
Etape 1
Etape 2
Etape 3
Etape 4

```
Appelation 0.1 أيْ أبا يا هايْ  أبّا آيْ هيّا
Mot-Liaison 0.1 ْ بِيهِمْ الحَاضِرْ إسْمُو مِنْها نْجِيبُوا خْوايِجْنا تْنالِي مَاكْ بِالأنْتِرْنَاتْ بَحْذَانا عَلِيهِمْ رَاكِبْ     خَالْةْ
Disfluences 0.1 ِي إنْتِيو إبنْ نْمِلْ بِش هَاةْ الحَذْ إةْ بَاسْ النْبِضْ الأوْرَ دِي بِي بَا بَسْ يَسْ وْفَا بِلْ نْفْ عَتْ الق
Ville 0.1 ِيرْ-القَصْعَةْ بُوعَرْقُوبْ لبُوعَرْقُوبْ لِبْرُويِسْ  لِبُوسَالِمْ بْرُويِسْ بُوعْرَادَةْ بُوسَالِمْ لغَارْ-دِمَاءْ  غَارْ-دِمَاءْ
Obligation 0.1 يِلْزْمِكْ  المُهِمْ لَابُدْ لَازْمُو يِلْزِمْ لَازْمِكْ لَازِمْ لَازْمْنِي يِلْزِمْنِي لَازِمْنِي
Ordre 0.1 الثَانِيَةْ
Choix 0.1 سِنُونْ أوْ  سِينَانْ  وِلَّى
Difference 0.1 فَرْقْ الفَرْقْ الدِيفِرُونْس
Distance 0.1 كِيلُومَاتِرْ
Place 0.1  البْقَايِعْ
Payement 0.1 لُو يَدْفَعْ يِدْفَعْ نْخَلْصُو بِالشَاكْ شَاكْ صَرِّفْ بِلْنَاتْ بِلْتْرِيتْ يِتْكَلَّفْلِكْ نْكَلْفْلِكْ خْلُصْ خَلَّصْ أصْرُفْ
Individu 0.1 الكُنْتْرُولُورْ الكُنْتْرُلُورْ صْحَابْهَا نَاسْ   بِالمُوَاطِنْ حَدْ النَاسْ
Future 0.1 بَاشْ
Possession 0.1 مْتَاعِي مْتَاعِكْ مْتَاعْهَا لِيلِكْ لِيلِي مْتَاعُو لِيكْ لِيَا  عَنْدْهَا لِيَّا مْتَاعْهُمْ مْتَاعِي مْتَاعْنَا
Satisfaction 0.1 سَايَايْ تْكِفِينِي  يْكَفِّينِي يِزِّينِي
Esperance 0.1 إنْشَاءْ-اللهْ إنْشَاءْ-اللهْ كَانْ-حَلَّانَا-القُدْرَةْ إنْ-شَاءْ-اللهْ إنْ-شَاءْ-اللهْ
Prix 0.1 ْ يِتْكَلِّفْ قَلْبْ-بْعَضْهُمْ سِعِرْ الفْرَانْوِي تَارِيفْ نْقَابِلِي يِطَلْعُولِي يِطَلْعِلِكْ جُمْلَةْ  الكَشْفْ لَبْرِي عَالسُومْ
Future 0.1 بَاشْ
Possession 0.1 مْتَاعِي مْتَاعِكْ مْتَاعْهَا لِيلِكْ لِيلِي مْتَاعُو لِيكْ لِيَا  عَنْدْهَا لِيَّا مْتَاعْهُمْ مْتَاعِي مْتَاعْنَا
Classement 0.1 ، تَالِتْ  التَانِي أخُرْ أخِرْ لُولَانِي   لُولَى التَالْتَةْ عَالأوَّلْ التَانِيَةْ لْخَرَانِيَّةْ الأخِيرْ تَانِي التَالِي أخَّر
Esperance 0.1 إنْشَاءْ-اللهْ إنْشَاءْ-اللهْ كَانْ-حَلَّانَا-القُدْرَةْ إنْ-شَاءْ-اللهْ إنْ-شَاءْ-اللهْ
Prix 0.1 ْ يِتْكَلِّفْ قَلْبْ-بْعَضْهُمْ سِعِرْ الفْرَانْوِي تَارِيفْ نْقَابِلِي يِطَلْعُولِي يِطَلْعِلِكْ جُمْلَةْ  الكَشْفْ لَبْرِي عَالسُومْ
Comparaison 0,1 وْ هَكَّكْ قُدَّامْ تَالِي قُدَّامِكْ القُدَّامْ بَعْدْهَا هَكَايَةْ مْتْعَدِّلِينْ أكْثَرْ بَعْدْهُمْ كِيمَ هَكَايْ كِيفْمَا نَفْسْ
```

Fig 2. Prepared corpus extract

- **Language model calculation:** In the language model learning corpus, including dialogue transcripts, all words (or semantic blocks) are replaced by class names. Finally, we use the SRILM[2] toolbox to learn language model including semantic classes. SRILM is a toolkit for building and applying statistical language models, primarily for use in speech recognition, statistical tagging and segmentation, and machine translation. The toolkit SRILM allows not only to build mainly n-grams language model but also to create models n-classes of words.

- **Evaluation of a language model: measure of perplexity:** Several measures are used to evaluate the quality of language model. We present perplexity as the most used method. Perplexity (PPL) is a quick method to evaluate the language models. It is commonly used for several years to judge the quality of a language model [14]. This evaluation metric is used to measure the prediction ability of a language model on a test corpus not seen during learning. The principle of perplexity is to check how much a language model can predict the word sequences of the language it is supposed to model. Perplexity is defined by:

$$\text{PPL}= 2^{-\frac{1}{n}\sum_{t=1}^{n} \log P(w_t|h)}$$

Where P $P(w_t|h)$ represents the probability proposed by the language model for the word w_t knowing the history h. The perplexity of our model is around 4.17. Indeed,

[2] http://www.speech.sri.com/projects/srilm/

the perplexity of a language model is between 1 and V, V is the size of vocabulary, that is to say the number of words that compose it. A reduced value of perplexity leads to better language model prediction capability. However, the value of perplexity alone does not mean much, it becomes useful when it is used to compare models with each other on the same test corpus. Hence, the model with the smallest perplexity is the best. As a result, we used the SRILM tool to construct an n-gram language model on the same training corpus and calculate its perplexity on the same test corpus in order to compare it with our n-class model. The table 4 below shows a comparison between the n-class model and the n-gram model in terms of perplexity.

Table 4. Value of perplexity calculated on the same test corpus

Type of model	Perplexity
n-gram	74.4641
n-classe	4.17

The perplexity of a language model permits this model to be evaluated as an isolated entity, regardless of its integration into the speech recognition system. As we have already mentioned, a low value of perplexity reflects a strong predictive power of language model. Thus, to better judge our choice of creating an n-class word model, we compared the created model with the 3-gram type model on the same test corpus of evaluation. Table 4 shows the very significant relative reduction in perplexity. These results are consistent with what could be expected: it is the classification based on semantic data that has minimized the perplexity of the language model obtained.

The value of the n-class model perplexity remains well below that of the 3-gram model on the test corpus. Interestingly, the same models as for the learning corpus have the lowest perplexity value on the test corpus. Thus, even if the obtained results are satisfactory, there are always successions of not observed classes in the learning for which the model will attribute a null probability. In fact, the estimation of probabilities depends on the size of the learning corpus.

4 Conclusion

The objective of this work is the construction of a statistical language model that is one of the components of automatic speech recognition system. In particular, we are interested in n-class language models by using semantic information for the creation of word classes. Indeed, we used the n-class model because it solves our problem of lack of Tunisian dialect data. Thus, since our field of work is limited “the Tunisian Railway Transport Network”, it may happen that some words are similar but they do not have the same probability of appearance, so their grouping in class will be possible.

The proposed method consists of three phases. The first represents the construction and processing phase of the corpus, which consists in building the corpus representing the domain to be studied and then transcribing and normalizing it. The second phase concerns the construction of language models. This phase groups together the steps of semantic tagging, word classification and language model calculation using the SRILM tool. The third phase concentrates on evaluating the constructed model by calculating its perplexity. In order to evaluate the model generated by our statistical language modeling system, we first calculated its perplexity rate, and in a second time we compared it to another model of constructed language by using the SRILM tool. The comparison is done on the same test data. On the one hand, the value of the perplexity of our n-class language model can be judged as weak, which necessarily reflects its satisfactory predictive power. On the other hand, the result of the comparison between the two models, n-class model and n-gram model, on the same evaluation corpus, proves that the predictive power of n-class model is better than that of n-gram model**,** which presents a high perplexity value.

References

1. Bahl L.R., Brown P.F., Souza P.V., Mercer R.L.: A tree-based Statistical Language Model for Natural Language Speech Recognition, IEEE Transactions on Acoustics, Speech and Signal Processing (1989).
2. Bigi B : Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction, Laboratoire Parole et Langage - Aix-en-Provence (2012).
3. Bilmes J. A., Kirchhoff K.: Factored Language Models and Generalized Parallel Backo, Proc. of Human Language Technologies, North American (2003).
4. Bouagres F. : Attelage de systèmes de transcription automatique de la parole, thèse de doctorat, université du Maine-Le Mans-France (2012).
5. Brown P. F., DellaPietra V. J., Souza P. V., Lai J. C., Mercer R. L.: Class-Based N-Gram Models of Natural Language, Computational Linguistics (1992).
6. Chelba C., Engle D., Jelinek F., Jimenez V., Khudanpur S., Mangu L., Printz H., Ristad E., Rosenfeld R., Stolcke A., Wu D.: Structure and Performance of a Dependency Language Model, Dans Proc. of the European Conf. on Speech Communication and Technology (Eurospeech) (1997).
7. Chelba C., Jelinek F.: Structured Language Modeling », Computer Speech and Language (2000).
8. Damnati G. : Modèles de langage et classification automatique pour la reconnaissance de la parole continue dans un contexte de dialogue oral homme-machine, thèse de doctorat, université d'Avignon et des pays du vaucluse (2000).
9. Farhat A., Isabelle J.F., O'Shaughnessy D.: Clustering Words for Statistical Language Models based on Contextual Word Similarity, Dans Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, USA (1996).
10. Jelinek F.: Continuous Speech Recognition by Statistical Methods, Proc. Of the IEEE (1976).
11. Ji G., Bilmes J.: Dialog Act Tagging Using Graphical Models, Proc. Of the IEEE Intl Conf. on Acoustics, Speech, and Signal Processing'ICASSP (2005).
12. Habash, N., Diab, M. and Rambow, O.: Conventional Orthography for dialectal Arabic,Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC-2012, (2012).
13. Kirchhoff K., Yang M.: Improved Language Modeling for Statistical Machine Translation, Proc. of the ACL Workshop on Building and Using Parallel Texts (ParaTex), Morristown, NJ, USA. Association for Computational Linguistics (2005).

14. Kneser R., Ney H.: Improved Clustering Techniques for Class-Based Statistical Language Modeling. Proc. European Conference on Speech Communication and Technology, Berlin, Allemagne (1993).
15. Lecorvé G. : Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole, thèse de doctorat, Université européenne de Bretagne (2010).
16. Masmoudi A., Ellouze M., Estève Y., HadrichBelguith L., Habash N.: A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition, LREC'2014, Reykjavik, Iceland (2014).
17. Masmoudi A Ellouze M., Estève Y., Bougares F HadrichBelguith L: Phonetic tool for the Tunisian Arabic, SLTU'2014, Russia, (2014).
18. Masmoudi, A., Habash, N., Khmekhem, M., Esteve, Y. and Belguith, L.: Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation, Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, p.608-619, (2015).
19. Masmoudi, A., Bougares,F., Ellouze, M., Estève, Y., Belguith, L.: Automatic speech recognition system for Tunisian dialect. Language Resources and Evaluation 52(1): 249-267 (2018).
20. Rosenfeld R.: Adaptive Statistical Language Modeling: A Maximum Entropy Approach, PhD Thesis, (1994).
21. Smaili K., Jamoussi S., Langlois D., Haton J.-P.: Statistical feature language model, Proc. of the 8th Intl Conf. on Spoken Language Processing'ICSLP, (2004).
22. Vergyri D., Kirchhoff K., Duh K., Stolcke A.: Morphology-Based Language Modeling for Arabic Speech Recognition, Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP), Jeju Island, South Korea, (2004).
23. Xu P., Chelba C., Jelinek F.: A study on richer syntactic dependencies for structured language modeling, Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL), Morristown, NJ, USA. Association for Computational Linguistics, (2002).
24. Zitouni I.: Linearly Interpolated Hierarchical N-gram Language Models for Speech Recognition Engines, IBM T.J. Watson Research Center, NY, Bell Labs Alcatel-Lucent, NJ,USA, (2008).
25. Zribi I., Boujelben R., Masmoudi A., Ellouze M., Belguith L., Habash N.: A conventionnal Orthography for Tunisian Arabic, LREC'2014, Reykjavik, Iceland, . (2014).