# Retour d'expérience : l'utilisation de l'apprentissage profond (deep learning) dans le contexte de l'analyse sémantique des langues peu dotées

Hammou Fadili[1, 2]
(1) Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
(2) Pôle Systèmes d'Information et du Numérique, Programme Maghreb de la FMSH Paris
190, avenue de France 75013, Paris, France
Hammou.fadili@cnam.fr / fadili@msh-paris.fr

## Résumé

On estime à plusieurs milliers le nombre de langues parlées dans le monde et seulement quelques dizaines disposent de ressources (informatiques, textuelles, etc.) permettant leur traitement automatique. Celles ne disposant pas ou disposant de peu de ressources sont appelées langues peu dotées (LPD). Plusieurs rapports de l'UNESCO affirment que la plupart des langues peu dotées sont en voie de disparition. De plus, plusieurs spécialistes des langues, estiment que leur disparition est accélérée par les phénomènes informatiques (internet, réseaux sociaux, etc.) qui les marginalisent encore plus. Cependant, d'après les mêmes spécialistes, l'intégration des langues peu dotées dans le monde des nouvelles technologies pourrait constituer une opportunité pour leur développement, leur sauvegarde et donc pour leur survie. En effet, mettre à disposition des utilisateurs des outils les incitant à la découverte et à la création dans les LPD, aidées par des passerelles avec d'autres langues mieux dotées (LMD), telles que les fonctionnalités : de liens, d'alignement, de traduction, d'analyse et de synthèse, etc. pourrait avoir un impact positif sur la popularité de leur utilisation et par conséquent sur leur développement. Dans cet article, nous présentons une expérience exploitant les nouvelles technologies d'apprentissage profond dans le contexte de l'analyse sémantique des langues peu dotées. Le but est de montrer à travers un exemple d'approche qu'on peut exploiter certaines technologies facilement adaptables aux langues soufrant du manque de ressources en termes de contenus et d'outils informatiques ; espérant que cela pourra, en plus, aider à sensibiliser et à inciter les chercheurs du domaine à proposer des solutions génériques intégrant dans leur conception le support des LPD.

## Abstract

**Feedback : use of deep learning in the context of poorly endowed languages.**
It is estimated that there are several thousand languages spoken in the world and only a few dozen have resources (tools, corpuses, annotations, etc.) for automatic processing. Those with little or no resources are called poorly endowed languages (PEL). Several UNESCO reports state that most

poorly endowed languages are endangered. In addition, several language specialists believe that their disappearance is accelerated by the phenomena of new technologies (Internet, social networks, etc.) which further marginalize them. However, according to the same specialists, the integration of poorly endowed languages into the world of new technologies could constitute an opportunity for their development, their preservation and therefore for their survival. Indeed, make available to the users, tools encouraging them to discover and to create in the PEL, helped by bridges with better endowed languages (BEL), such as functionalities of: links, alignment, Translation, analysis, etc. Could have a positive impact on the popularity of their use and consequently on their development. In this article, we present an experiment exploiting the new technologies of deep learning in the context of the semantic analysis of PEL. The aim is to show through an example of approach that we can exploit certain technologies that are easily adaptable to languages suffering from the lack of resources in terms of content and computer tools; Hoping that it will also help to raise awareness and encourage researchers in the field to propose generic solutions integrating in their design the support of LPD.

MOTS-CLÉS : Langues peu dotées, apprentissage profond, modèles de langue, sémantique, représentations vectorielles des mots.

KEYWORDS: Poorly endowed languages, deep learning, language model, semantics, word embeddings.

# 1 Introduction

On estime à plusieurs milliers le nombre de langues parlées dans le monde et seulement quelques dizaines disposent de ressources (informatiques, textuelles, etc.) permettant leur traitement automatique. Celles ne disposant pas ou disposant de peu de ressources sont appelées langues peu dotées (LPD). Plusieurs rapports de l'UNESCO affirment que la plupart des langues peu dotées sont en voie de disparition. De plus, plusieurs spécialistes des langues, estiment que leur disparition est accélérée par les phénomènes informatiques (internet, réseaux sociaux, etc.) qui les marginalisent encore plus. Cependant, d'après les mêmes spécialistes, l'intégration des langues peu dotées dans le monde des nouvelles technologies pourrait constituer une opportunité pour leur développement, leur sauvegarde et donc pour leur survie. En effet, mettre à disposition des utilisateurs des outils les incitant à la découverte et à la création dans les LPD, aidées par des passerelles avec d'autres langues mieux dotées (LMD), telles que les fonctionnalités : de liens, d'alignement, de traduction, d'analyse et de synthèse, etc. pourrait avoir un impact positif sur la popularité de leur utilisation et par conséquent sur leur développement.

C'est dans ce sens que nous avons mené une expérience exploitant les nouvelles technologies d'apprentissage profond pour le traitement automatique d'un cas d'une langue peu dotée. Nous avons essayé à travers ce travail de privilégier des méthodes basées sur des apprentissages non/peu supervisées et des méthodes statistiques capables d'effectuer des traitements sur des données brutes n'ayant subi aucun traitement au préalable. Le but est de montrer la faisabilité de contourner les

problèmes liés aux manques de données structurées, annotées, etc., d'outils et de règles de traitements, dont souffrent les LPD. Nous espérons, en plus, à travers ces expériences, aider à sensibiliser la communauté travaillant dans le domaine du TALN en général, et les inciter à proposer des approches et outils génériques, réutilisables et applicables dans le contexte de n'importe quelle langue y compris celles peu dotées. Cet article est organisé comme suivant, la première partie est consacrée à un bref rappel sur les langues peu dotée. La deuxième partie rappelle quelques éléments technologiques retenus en termes de modèle de langue, de modèles de données et de traitements. La partie suivante est consacrée à la description de l'ensemble des éléments au sein d'une architecture adaptée, à leur fonctionnement et aux résultats des tests menés. La dernière partie conclue le présent article.

## 2 Contexte des langues peu dotées

D'une manière générale, les langues peu dotées, sont des langues qui souffrent de plusieurs problèmes : problèmes liés à la graphie, au manque d'un système d'écriture stable, au manque de ressources informatiques et linguistiques. Le manque de ressources langagières concerne les dictionnaires, thésaurus, corpus traités, etc. ; le manque d'outils numériques concerne les outils du traitement automatique de la langue naturelle : analyseurs morphologiques, syntaxiques, sémantique, etc. Tous ces éléments rendant difficile, voire impossible l'analyse sémantique des langues peu dotées, peuvent être classés suivant les 3 axes ci-après :

Modélisation & modèle de langue : la langue naturelle est très complexe en général. Ceci est dû au nombre important (presque infini) de cas possibles d'utilisation, d'exceptions et de règles, etc. La modélisation des caractéristiques de la langue naturelle est une grande problématique, d'une manière générale ; problématique encore plus accentuée dans le cas des LPD.

Connaissances & données : un autre problème dont souffrent les LPD concerne les données prétraitées qui sont d'une grande nécessité dans les systèmes de traitements des données. Elles sont utilisées pour instancier les modèles des données et pour exprimer les règles de raisonnement et d'apprentissage. Le manque de ce type de données, dans le cas des LPD constitue un frein majeur pour leur exploitation automatique.

Outils : les outils informatiques nécessaires à l'automatisation des tâches des traitements, bien adaptés au LPD sont également rares.

Afin de contourner ces problèmes, nous avons expérimenté et réutilisé une solution ayant fait ses preuves dans le cas d'autres langues mieux dotées (l'Anglais et le Français,). Elle a l'avantage d'être indépendante ou plutôt peu dépendante de la langue traitée.

# 3 Démarche expérimentale

Cette partie a pour but de décrire l'expérience menée, en étudiant le comportement sur une LPD, en l'occurrence le Berbère, de l'approche mise en place pour le Français et l'Anglais. Cette approche étant générique, i.e. pas ou peu dépendante de la langue traitée, nous a permis de contourner certains problèmes liés aux LPD, décrits précédemment. On y exploite des apprentissages non supervisés ou peu supervisés des modèles de représentation et de traitements pour le traitement automatique de la sémantique du texte comme la désambiguïsation.

## 3.1 Modèle de langue

Pour le choix et la représentation du modèle de langue, nous avons besoin de résoudre et de contourner certaines difficultés : celles relatives à la modélisation et à la formalisation de la langue naturelle, puis celles relatives aux choix technologiques pour les représenter et aux réalisations informatiques pouvant les supporter. Dans le premier cas, nous avons à modéliser deux notions importantes du domaine de l'analyse des données non structurées, à savoir la notion de contexte et la notion des relations sémantiques afin de mieux caractériser la sémantique. Ces notions déjà modélisées pour d'autres LMD ont été validées et adaptées pour le cas du Berbère ; grâce au concourt des linguistiques du domaine qui nous ont aidé à adapter le modèle et créer un modèle générique de langue.

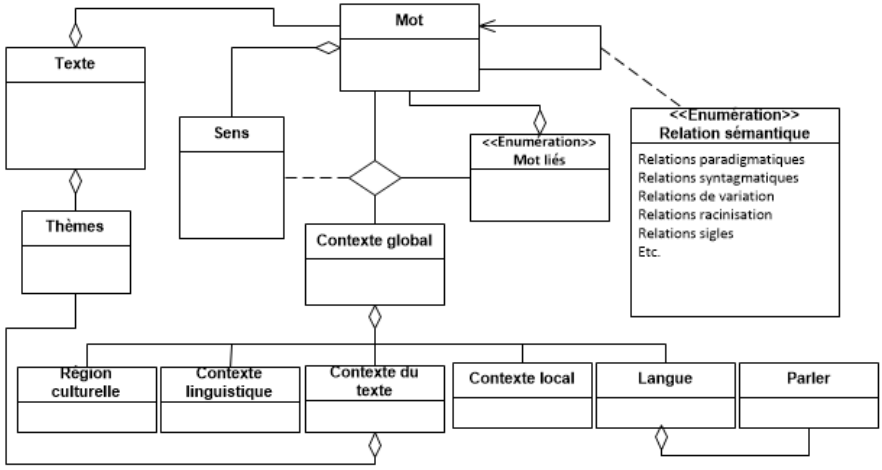


FIGURE 1: Modèle pour l'apprentissage

Sur le plan technologique, plusieurs problèmes devaient être également pris en compte et résolus. Cela concerne principalement les choix technologiques et applicatifs pouvant d'une part supporter le modèle de langue et d'autre part optimiser et simplifier les traitements. Nous avons évité

d'utiliser les modèles technologiques de langue basés sur des représentations vectorielles « sparses » de type sac de mots classique, représentant les mots comme des vecteurs dans des espaces vectoriels de très grandes dimensions (taille de tout le vocabulaire) ; difficile à mettre en place. Pour cela, nous avons fait le choix d'adopter les nouvelles représentations vectorielles « denses » de type « Word embeddings » et son implémentation « word2vec ». Pour cela, nous avons utilisé, les 2 implémentations de Word2vec : Skip-gram (Mikolov et al. 2013a) et CBOW (Mikolov et al. 2013b). Le but est d'entrainer un réseau de neurones profond pour obtenir une représentation vectorielle sémantique réduite de chaque mot à partir de sa représentation initiale et ses contextes locaux.

Dans le cas de Skip-Gram :

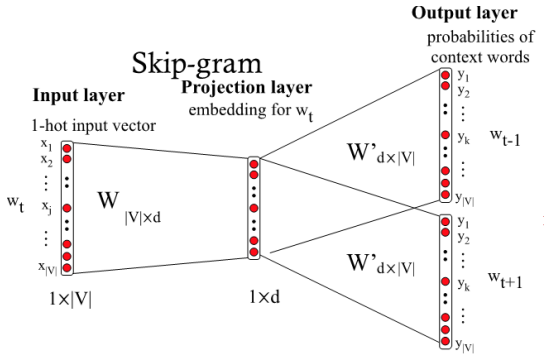


FIGURE 2: Skip-gram

L'apprentissage se fait en maximisant la fonction objective des Log probabilités.

La normalisation des probabilités se fait par un Softmax.

$$\Sigma_{t=1}^{T}\Sigma_{j=-c,j\neq 0}^{c}\log P(w^{t+j}\mid w^{t})$$

$$P(w^{t+j}\mid w^{t})=\frac{\exp(\mathbf{v}_{w^{t+j}}\cdot\mathbf{v}_{w^{t}})}{\Sigma_{i=1}^{V}\exp(\mathbf{v}_{w_i}\cdot\mathbf{v}_{w^{t}})}$$

k est la taille du contexte local

Dans le cas de CBOW (Continuous Bag Of Words) :

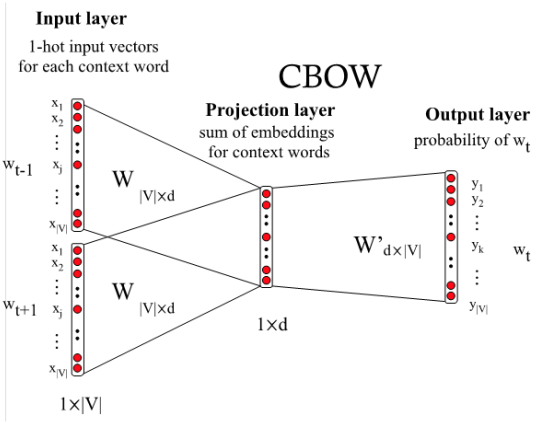


FIGURE 3: CBOW

L'apprentissage se fait en maximisant la fonction objective des Log probabilités.

La normalisation des probabilités se fait par un Softmax.

$$\Sigma_{t=1}^{T} \log P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c}).$$

$$P(w^t \mid w^{t-c},...,w^{t-1},w^{t+1},...,w^{t+c}) = \frac{\exp(\mathbf{v}_{\bar{w}^t} \cdot \mathbf{v}_{w^t})}{\Sigma_{i=1}^{V} \exp(\mathbf{v}_{\bar{w}^t} \cdot \mathbf{v}_{w_i})}$$

$\mathbf{v}_{\bar{w}^t}$ la moyenne (pondérée) des représentations vectorielles des mots contextuels de $w^t$

Cette technologie permet de coder l'historique des utilisations d'un mot dans un vecteur dense d'environ 300 dimensions. Il s'agit d'apprendre une projection d'un espace vectoriel initial « sparce » dans un espace vectoriel « sémantique » et « dense » représentant toutes utilisations passées afin de prédire les utilisations futures. Cette technologie a été testée avec succès, bien adaptée pour toutes les langues. Dans notre cas, la technologie « Word embeddings » a été exploitée pour instancier et représenter notre modèle sémantique d'apprentissage. Cela correspond à une extension et un enrichissement de la technologie « Word embeddings » initiale. Ces travaux nous ont permis, d'une part, une formalisation approfondie et générique et une prise en compte de certaines caractéristiques importantes de la langue naturelle, pouvant influencer le sens ; et d'autre part, l'exploitation de certaines technologies de pointe sur lesquelles nous avons apporté des améliorations, afin de faire progresser certains résultats dans ce domaine en général et dans le contexte des LPD en particulier. En l'appliquant à un cas concret d'une langue, nous avons pu, montrer l'apport considérable de ces technologies pour le traitement automatique des LPD.

## 3.2 Génération d'instances pour l'apprentissage

Cette partie est consacrée à la présentation des approches statistiques basées sur des apprentissages non supervisés pour l'extraction des caractéristiques sémantiques et des thèmes latents du texte. Ceci afin de représenter convenablement les mots du texte et aussi pour compléter l'instanciation du modèle sémantique étendu de langue pour l'apprentissage.

Pour cela, nous avons exploité 3 types d'outils et de méthodes :

- Modèles neuronaux
    - Représenter vectoriellement les mots
    - Word2vec - Skip-grams, CBOW (Mikolov, 2013)
- Outils de TALN
    - Extraction du contexte local
    - Extraction du contexte linguistique
- Modèles de groupements thématiques
    - Extraire les thèmes latents
    - Latent Dirichlet allocation - LDA (Blei, 2009)

Des éléments du contexte tels que la langue, le parler, etc., sont instanciés à la main une seule fois avant l'application du processus. Automatiser cette tâche est possible, car des travaux sur la détection automatique de la langue et des parlers berbères existent, on peut citer par exemple ceux de (W. Adouane et al., 2016).

Dans le cas des LPD ne disposant pas de Wordnet ou équivalent, nous avons adopté, contrairement à l'approche initiale, une méthode de « groupement » ou « Clustering » non supervisée pour le calcul des classes des sens. Nous avons testé et exploité la méthode de (Schütze, 1998), qui est une méthode de désambiguïsation automatique et non-supervisée basée sur les résultats d'une « clustérisation » des données et qui prend en compte les cooccurrences du second ordre dans la représentation du contexte des mots.

La combinaison de toutes les caractéristiques permet la représentation vectorielle du modèle sémantique pour l'apprentissage (contexte local, contexte linguistique, thématiques et domaines traités (contexte global)).

### 3.3 Outils

Sur le plan des outils, plusieurs plateformes (Weka, Rapidminer, Orange, ...) et classifieurs (Bayes, Arbres de décision, SVM, Réseaux de neurones…) ont été testés avant de faire le choix d'exploiter un réseau de neurones profond pour le Traitement Automatique de la Langue Naturelle (TALN). Il suffit que l'UNICODE soit supporté, pour pouvoir exploiter ces plateformes et ces technologies, ce qui est le cas du berbère.

## 4 Architecture

Nous avons encapsulé un certain nombre d'éléments, décrits dans ce document, dans un processus d'analyse sophistiqué permettant l'extraction automatique des caractéristiques sémantiques du modèle sémantique retenu, comme source d'entrée pour un réseau de neurone profond. Ceci afin de faciliter le processus d'analyse, d'interprétation et d'exploitation sémantiques automatiques des données non structurées.

### 4.1 Apprentissage profond (Deep learning) pour le TALN des LPD

Notre contribution initiale, en plus d'être basée sur des recherches améliorant les modèles de langues existantes, permettant de caractériser au maximum les textes, leurs mots, leurs utilisations et leurs sens, et par conséquent permettant d'améliorer leurs analyses et leurs interprétations ; est basée sur l'apprentissage profond des modèles étendus de représentation et de traitement de la sémantique des textes. Ces architectures profondes sont des réseaux comportant plusieurs couches, pouvant modéliser avec un haut niveau d'abstraction des modèles de données, articulés autour de transformations non linéaires. Dans cette partie, nous avons exploité la modularité des architectures profondes pour mieux les adapter au contexte des LPD. Par exemple, l'emplacement des modèles symboliques dans le processus général, était problématique ; car ils sont difficiles à mettre en place, surtout dans le cas des langues peu dotées. Un texte doit y être représenté par les mots le constituant ainsi que par les propriétés issues des différents traitements linguistiques ou des formalisations spécifiques telles les ontologies, faisant défaut dans le cas des LPD (cf. paragraphe ci-dessus).

Pour contourner ces difficultés, nous avons privilégié l'intégration des modèles numériques statistiques en amont du processus. Ces modèles sont basés sur des calculs mathématiques dans des espaces sémantiques optimisés, bien adaptés. L'avantage est qu'on n'a pas besoin d'informations (pré)-traitées au préalable pour modéliser et décrire ce type de modèles ; on y exploite directement les données à analyser et les modèles mathématiques pour déduire les modèles de représentations.

### 4.2 Architecture

Concrètement, les tests ont été réalisés grâce à un réseau de neurones multicouches, permettant des apprentissages mixtes comme suivant :

- les premières couches ou couches basses ont été consacrées à l'extraction des informations latentes (features). Elles utilisent des apprentissages non supervisés, exploitant des approches numériques basées sur les fréquences et les agencements des mots, et sur les modèles mathématiques statistiques et probabilistes, pour le codage de la sémantique

- les dernières couches ont été consacrées à la prise de décision. Elles utilisent des apprentissages supervisées, exploitant des approches symboliques basées sur les structures des données, leur description ainsi que sur des systèmes de règles, générées dans les couches précédentes, pour effectuer les traitements

Cette démarche a l'avantage de combiner les deux approches, à des degrés différents, et avec des ordonnancements spécifiques. Son originalité réside dans le fait que, d'une part, elle intègre des notions importantes dans le modèle de langue : la notion de contexte global et la notion des relations sémantiques ; d'autre part, elle exploite ces notions pour générer un modèle sémantique riche de données sur lequel on applique des algorithmes d'apprentissage.

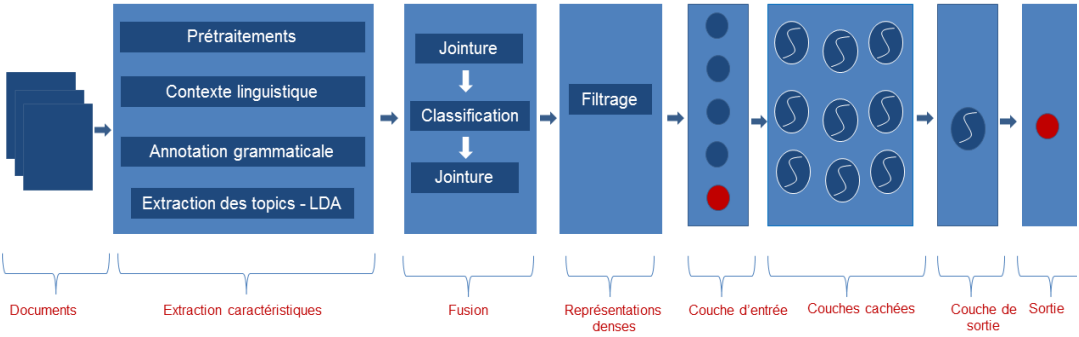


FIGURE 4: Réseaux de neurones multicouches

## 5 Tests

Nous avons développé et testé des modules permettant d'extraire les caractéristiques sémantiques et les instances dans le format « .arff » supporté par la plateforme d'apprentissage Weka. Nous avons séparé les données générées en trois parties comme suivant :

- 20% des données pour la validation, ceci afin d'optimiser les hyper-paramètres du système : le pas d'apprentissage, le type de la fonction d'activation et le nombre de couches

- 60% pour l'entrainement, ceci afin d'estimer les meilleurs coefficients (wi) de la fonction du réseau de neurones, minimisant l'erreur entre les sorties réelles et les sorties désirées

- et 20% pour les tests, ceci afin d'évaluer les performances du système.

La génération du fichier .arff, s'est fait grâce, dans un premier, aux modules développés, qui nous ont permis d'instancier le modèle (POS, Contexte local, Topics, classe, etc.) via un fichier .csv. La plateforme Weka a ensuite été utilisée pour l'implémentation d'un réseau de neurones de type « perceptron multicouche » appliqué aux instances pour l'évaluation.

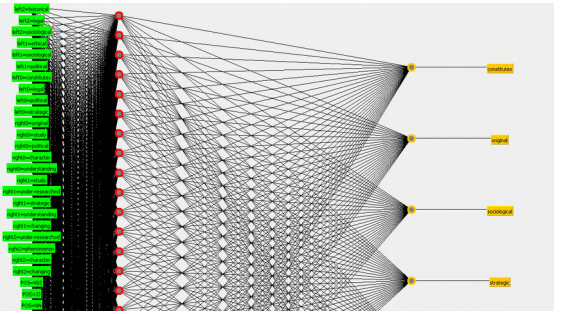

FIGURE 5: Application du perceptron multicouche

Ci-après quelques résultats des tests :

| | |
|---|---|
| Correctly Classified Instances | 71.4286% |
| Incorrectly Classified Instances | 28.5714% |

TABLE 6 : Vue d'ensemble

| | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0.917 | iyya |
| | 1 | 1 | 1 | 1 | anezli |
| | 1 | 1 | 1 | 1 | tasnametti |
| | 1 | 1 | 1 | 1 | astrategic |
| | 0.5 | 1 | 0.667 | 0.917 | tisfrass |
| | 0.5 | 1 | 0.667 | 0.917 | yann |
| | 0 | 0 | 0 | 0.917 | azamul |
| Weighted Avg. | 0.571 | 0.714 | 0.619 | 0.952 | |

TABLE 7 : Mesures

| a | b | c | d | e | f | g | <-- classified as |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | \| a=iyya |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | \| b= anezli |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | \| c= tasnametti |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | \| d=astrategic |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | \| e=tisfras |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | \| f=yann |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | \| g= azamul |

TABLE 8 : Matrice de confusion

Une fois le modèle d'apprentissage correctement instancié, les résultats des tests montrent des performances similaires que sur les langues mieux dotées. Le système à réussit grâce à l'apprentissage sur une partie des instances à déduire le sens réel de tous les mots qui étaient mal classés au départ.

# 6 Conclusion

Cet article décrit une expérience basée sur une approche d'apprentissage indépendante de la langue traitée, pour instancier le modèle de données d'apprentissage et lui appliquer des méthodes d'apprentissages non ou peu supervisés. Elle a l'avantage de contourner les problèmes majeurs rencontrés dans l'analyse des données non structurées dans le contexte des langues peu dotées, à savoir le manque d'outils et de données annotées, sémantiquement et automatiquement exploitables. Les expériences menées sur des exemples confirment d'une part l'apport considérable du modèle proposé pour la détection du sens réel des mots dans le texte et d'autre part, l'application de l'approche sur les langues peu dotées. Nous espérant que cette expérience peut aider à sensibiliser et à inciter les chercheurs du domaine à concevoir et à développer des solutions génériques applicables à plusieurs langues dont les LPD. Ce qui pourrait contribuer à développer et peut-être même à sauvegarder ce type de langue en voie de disparition, faute de moyens et soutiens.

# Références

ADOUANE W., SEMMAR N., JOHANSSON R. (2016). Romanized Berber and Romanized Arabic Automatic Language Identification Using Machine Learning. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,* 53–61, Osaka, Japan.

BLEI D., LAFFERTY J. (2009). Topic Models. *Text Mining, Classification, Clustering, and Applications* A. Srivastava and M. Sahami, editors. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Deerwester S. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.

Fadili H. (2013). Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexico-semantic relations and the concept of the context. *IEEE-AICCSA,* Ifrane, (Morroco), May.

Mikolov T., Chen K.,Corrado G., Dean J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint,* arXiv:1301.3781.

Mikolov T., Sutskever I., Chen K.,Corrado G., Dean J. (2013b).Distributed representations of phrases and their compositionality. In *NIPS*.

Schütze H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97–124.

Turney D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.