Outils, méthodes et problèmes de traitement d'un corpus multilingue: le cas du journal "Simpaticuni" (Tunis, 1911-1933)

Mériem Zlitni

Outils, méthodes et problèmes de traitement d'un corpus multilingue : Le cas du journal *Simpaticuni* (Tunis, 1911-1933)

Mériem Zlitni
MoDyCo, Université Paris Ouest Nanterre, 92000 Nanterre
meriem-zlitni@gmx.fr

RESUME

Dans cette communication nous voudrions présenter les outils, méthodes et problèmes de traitement d'un corpus multilingue. Un des objectifs de notre thèse est d'identifier et d'analyser le tissu linguistique du *Simpaticuni*, journal italien, prétendu dialectal, édité à Tunis de 1911 à 1933. Les textes composant le corpus sont des chroniques rédigées en langue hybride, mélangeant le sicilien et l'italien normé avec des occurrences de la variété dialectale d'arabe tunisien et du français. Nous voudrions examiner plus particulièrement le fonctionnement de l'emprunt à l'arabe tunisien. Nous proposons d'exposer les problèmes techniques auxquels nous sommes confrontés tels que le grand format des feuilles du journal et la variation linguistique qui rendent la numérisation difficile. Nous évoquerons le problème de l'étendue du corpus et de sa représentativité. Nous présenterons les logiciels que nous prévoyons d'employer tels que *Lexico3* et *Unitex*, ainsi que les méthodes envisagées.

ABSTRACT

Tools, methods and issues to process a multilingual corpus : The case of the journal *Simpaticuni* (1911-1933)

In this paper, we present the tools, methods and problems to process a multilingual corpus. One objective of this thesis is to identify and analyze the linguistic fabric of *Simpaticuni*, an Italian newspaper, claimed to be dialectal, that was published in Tunis from 1911 to 1933. The texts of this corpus are chronicles written in a hybrid language that mixes Sicilian and normalized Italian with some occurrences of various dialects of Arabic and French. We will examine more closely how Arabic Tunisian has been used. We will also explain some of the technical issues we face, such as newspapers' large format sheets and linguistic variation, which makes scanning difficult. We will address the issue of the extent of the corpus and its representativeness. We will present the softwares we plan to use – namely *Lexico3* and *Unitex*, as well as proposed methods.

MOTS-CLES : Variations linguistiques – Représentativité du corpus – Numérisation – Lexico3 – Unitex.
KEYWORDS : Linguistic variations – Representativeness of the corpus – Scanning – Lexico3 – Unitex.

1 Présentation du corpus

Notre corpus a été constitué à partir du dépouillement du *Simpaticuni*, journal italien, prétendu dialectal[1], édité à Tunis de l'année 1911 (25 juin, n°1) à l'année 1933 (9 septembre, n°1103)[2]. Le

[1] Le sous-titre du journal a été modifié à plusieurs reprises. Néanmoins, entre l'année 1912 (n°52) et l'année 1928 (n°896), le sous-titre *Politico, Letterario, Umoristico, Dialettale* (litt. politique, littéraire, humoristique, dialectal) a été adopté de façon constante.

[2] Le journal *Simpaticuni* se caractérisait par une singularité du fond et de la forme. Il représentait une exception dans le paysage journalistique de la Tunisie coloniale dont l'effervescence culturelle s'est traduite par la publication d'environ 12 3 titres de journaux italiens (Brondino, 1998).

journal est principalement conservé aux Archives Nationales de Tunisie sur support papier.

Le travail de dépouillement ainsi que la description technique du contenu de quelques échantillons choisis avec un écart de temps de dix ans auraient révélé l'emploi de plusieurs variétés et niveaux de langues.

De manière générale, les rubriques proposées dans les colonnes du journal sont rédigées en langue italienne ou bien en dialecte sicilien. Toutefois, nous avons observé la régularité dans la publication d'un genre journalistique dont les auteurs employaient une langue métissée.

Il s'agit d'une chronique[3] dans laquelle ses divers signataires[4] proposent une description anecdotique de la vie quotidienne de la communauté italienne de Tunisie en utilisant un idiome hybride, présentant un mélange entre sicilien et italien normé avec des occurrences de la variété dialectale d'arabe tunisien et du français.

Cette chronique se caractérise par la diversité de ses auteurs, ce qui pourrait se traduire par une hétérogénéité énonciative et, par conséquent, par la présence de plusieurs idiolectes.

Ainsi, nous avons délimité notre corpus en réalisant un choix rigoureux des textes et en nous basant sur divers critères.

En premier lieu, nous avons tenu compte de la récurrence presque systématique des chroniques dans les numéros du journal.

Nous avons pris en considération le genre journalistique des unités discursives dans le but de choisir un corpus homogène[5]. Enfin, ce choix a également été dicté par la singularité linguistique du contenu de cette ressource textuelle (Bilger, 2000 ; Moirand, 2007).

Nous précisons aussi l'intérêt sociolinguistique et historique de ces textes d'oral-transcrit qui représenteraient à ce jour un des rares témoignages de cette langue particulière employée par les locuteurs d'origine sicilienne (Somai, 2000b).

Dans la partie qui suit, nous proposons un bref aperçu de la situation sociolinguistique de la communauté italienne installée dans la Tunisie coloniale afin de mieux comprendre le contexte historique et linguistique de l'époque.

2 Cadre sociolinguistique

D'un point de vue historique, notre corpus (1911-1933) est contemporain du protectorat français en Tunisie (1881-1956). Pendant cette période, une importante communauté d'origine italienne se renforça.

Entre 1815 et 1861, des réfugiés politiques fuyant les violences précédant l'Unité italienne, et des juifs Livournais exerçant le commerce s'installent sur le territoire tunisien. Ce groupe était instruit et aisé matériellement (Davi, 2000 : 101).

Entre 1861 et 1930 approximativement, un nombre important d'Italiens d'origine essentiellement sicilienne, constituant un prolétariat majoritairement pauvre et illettré, immigre en Tunisie où

[3] Le titre et le sous-titre de ce type de chronique pouvaient présenter des variations d'un numéro à un autre. En voici quelques exemples : *Al Caffè del Casino* (*Sceni di lu veru*), n°4 (5-6 août 1911, p.1) ; *Carruzzedda sfumata* (*Sceni successi pri daveru*), n°34 (21 juillet 1912, p.1) ; *Li iocu di lottu* (*Sceni chi succerinu*), n°602 (5 mai 1923, p.1), etc.

[4] Nous avons répertorié plusieurs signataires. Néanmoins, certains noms d'auteurs sont plus fréquents que d'autres tels que : *R. C.,* abréviation de *Rosario Cunsolo* qui s'occupa de la direction du journal de 1912 jusqu'en 1933 ; *Mastru 'Mbrogghia* ; *V. Mongelli* ou *V. M.* ; *Marco Visconti* ou *M. V.* ; *Il Figaro* ; *Briscula ; A Canzunara* ; *Viri a tutti* ; *Lu Stigghiolu ; Don Cocò* ; etc.

[5] Selon F. Rastier (2011 : 71-72), « […] Pour parvenir à des traitements automatiques spécifiques et efficaces de corpus, il convient en effet de tenir compte des genres, pour adapter les stratégies d'interrogation et de traitement. La stipulation préalable des genres permet de simplifier les traitements, ne serait-ce qu'en éliminant les ambiguïtés […] ».

s'offrent de plus larges possibilités d'emploi grâce aux travaux publics amorcés par le gouvernement du protectorat.

Ainsi, ce nouveau flux migratoire en provenance de l'Italie méridionale a considérablement modifié la physionomie et la consistance numérique de ce groupe ethnique. En effet, l'élément sicilien représentait désormais plus de 75% de l'ensemble de la communauté italienne (Loth, 1905: 106).

En ce qui concerne les langues parlées, les Italiens de Tunisie formaient un groupe non homogène dont les locuteurs provenaient de diverses régions d'Italie et appartenaient à des classes socioculturelles distinctes.

Cette scission était également perceptible sur le plan linguistique. En effet, il semble que la bourgeoisie, formée de juifs Livournais, était trilingue puisqu'elle s'exprimait à l'écrit comme à l'oral en langue italienne, parlait le français et l'arabe (Cohen, 1964 : 11).

En revanche, les migrants siciliens, majoritairement analphabètes et dialectophones[6], ne possédaient aucune connaissance de leur langue nationale[7]. Ainsi, il semble que le répertoire linguistique de ce groupe était composé essentiellement du dialecte sicilien, ou plus précisément de l'un des dialectes parlés en Sicile selon la province d'origine.

Dans les premières décennies qui ont suivi son installation, le langage oral de la collectivité sicilienne de Tunisie aurait subi deux transformations majeures en perdant ainsi toute ressemblance avec les parlers des Siciliens d'Italie.

La première transformation consisterait en l'uniformisation des différents parlers par le phénomène de la *koinéisation* (Bartens, 2000: 9), en d'autres termes la naissance d'une variété dialectale commune ou *koinè* conséquemment à l'augmentation de la communication intercommunautaire.

D'après Pendola (2000a: 14; 2000b: 84), le parler tunisien aurait servi de vecteur pour l'uniformisation de la langue de la communauté sicilienne. Il semblerait que lorsque deux termes d'origine sicilienne indiquaient le même objet, ils étaient abandonnés et le terme correspondant au dialecte tunisien était emprunté[8].

En second, la *koinè* se serait transformée en une langue de contact composite mélangeant le sicilien, l'arabe dialectal, l'italien et le français.

Cette langue commune de l'oralité, qui, selon certains témoignages, était comprise par la population multiethnique et plurilingue de la Tunisie, se serait développée afin de répondre spontanément à un besoin quotidien d'intercommunication entre des groupes de langues maternelles différentes[9].

Ainsi, il est possible de supposer que la situation socioéconomique modeste et les petits métiers (pêche, artisanat, agriculture, commerce, etc.) que les Siciliens exerçaient auraient permis l'établissement d'un contact direct avec le prolétariat tunisien qui partageait les mêmes conditions précaires.

Ces affinités auraient contribué à la naissance de relations de proximité (voisinages, rapports avec les commerçants, rencontres familiales et associatives) propices aux phénomènes d'interférences

---

[6] T. De Mauro (2005 : 57-59) souligne que l'émigration italienne a particulièrement touché les régions où le pourcentage d'illettrisme s'élevait à plus de 75%. Selon le linguiste, les migrants italiens étaient en grande partie analphabètes et, par conséquent, monolingues. Ce phénomène était également perceptible en Tunisie puisque les migrants d'origine sicilienne étaient en majorité illettrés à leur arrivée.

[7] Il semble que les locuteurs siciliens aient quitté l'Italie à une époque où la scolarisation n'était obligatoire que pour deux ans selon la *Legge Coppino* de 1877 (Alfieri, 1992 : 835).

[8] M. Pendola (2000a : 14 ; 2000b : 84) donne l'exemple du mot *abricot* lequel possède plusieurs désignations dans les divers dialectes de Sicile : *varcocu*, *bbarcocu*, *piricoculu*, *pricocu.* Ces différents termes auraient été remplacés par l'équivalent en arabe dialectal *mešmāš* que les Siciliens de Tunisie ont adopté sous la forme *musce mesce*.

[9] Selon G. Berruto (2002 : 183-184), cette variété d'italien, fortement mélangée à la langue du pays d'accueil, subit généralement des interférences au niveau lexical et beaucoup moins en morphosyntaxe. Le linguiste parle dans ce cas de *varietà rilessicalizzata*, soit une langue ayant gardé sa base phonologique, morphologique et syntaxique mais possédant un lexique varié emprunté en grande partie à la langue parlée sur place. Sur le langage hybride employé par la communauté sicilienne de Tunisie, consulter M. Pendola (2000a ; 2000b).

linguistiques.

Ce climat de symbiose serait à l'origine de la transformation de la koinè employée par les locuteurs siciliens en une langue composite présentant un mélange entre le parler sicilien, l'italien, l'arabe tunisien et le français.

En ce qui concerne la situation sociolinguistique des locuteurs tunisiens avant le protectorat, D.-E. Kouloughli (2007 : 100-106) affirme que celle-ci était caractérisée par l'emploi du turc en tant que langue officielle de l'administration.

La variété littéraire de l'arabe, que seule une minorité d'arabophones instruits connaissait, était réservée à l'écrit, et à un usage religieux et juridique.

Sinon, l'ensemble de la population, sans distinction de classe sociale ou de niveau d'instruction, employait dans tout type de communication orale le dialecte tunisien.

Durant l'époque coloniale, la langue turque disparaîtra du paysage linguistique et sera remplacée par le français, langue du colonisateur, qui va cumuler les fonctions culturelles et techniques, ne laissant à l'arabe classique qu'une fonction essentiellement littéraire et religieuse. Quant à l'arabe parlé, il sera cantonné à des usages quotidiens, intimes et informels.

3 Problématique et objectifs de la thèse

La problématique envisagée dans notre thèse concerne les aspects linguistiques et extralinguistiques relatifs aux phénomènes de contacts entre locuteurs arabophones et locuteurs appartenant à la communauté italienne installée en Tunisie depuis le XIX^{e} siècle.

A. Lakhdhar (2006), M. Pendola (2000a ; 2000c) et A. Somai (2000a ; 2000b) se sont intéressés à la question des échanges linguistiques dans la Tunisie coloniale, en s'appuyant notamment sur des exemples prélevés dans des numéros du journal *Simpaticuni*.

Fondées sur des relevés ponctuels, ces études ont été menées essentiellement sur des lexèmes. Elles n'ont pas donné lieu à une analyse lexicologique exhaustive et n'ont pas considéré le tissu phrastique dans lequel ils s'insèrent.

Nous voulons ainsi combler une partie de cette lacune en ayant notamment recours à des outils et logiciels afin de réaliser une analyse linguistique fine et originale.

Un des objectifs de cette recherche est d'analyser le tissu linguistique des chroniques du journal *Simpaticuni* afin d'identifier les particularités phonologiques, morphosyntaxiques, sémantiques et lexicologiques de la langue employée.

Nous voulons établir ainsi le degré de sicilianité du texte, en déterminant les traits siciliens, méridionaux et italiens (Grassi *et al.*, 2005). L'intérêt de cette approche est l'identification de l'idiome utilisé dans ce type de chronique. La question de savoir si nous sommes en présence d'un parler sicilien, d'un langage régional, ou bien d'une langue hybride se pose.

Nous sommes également confrontés à l'usage alterné de formes dialectales, méridionales ou bien siciliennes, et de formes plus ou moins italianisées d'un même mot dans le même texte. L'objectif est d'expliquer les conditions de ce type de variation linguistique.

A titre d'exemple, dans l'un des textes du corpus[10], nous avons relevé un cas de consonne cacuminale[11],

[10] Il s'agit de la chronique intitulée *Doppu lu futtiballi (Sceni di lu veru)* publiée dans le *Simpaticuni* n°658 de l'année 1924 (31 mai, p.2) et signée *Schut*.

[11] La consonne cacuminale, phénomène de phonétique articulatoire typique des régions situées à l'extrême sud de l'Italie (Calabre, Salento), est largement attesté en Sicile où l'occlusive voisée [d] est prononcée avec plus ou moins d'énergie, simple ou géminée, suivant les variantes locales (Grassi *et al.*, 2005 : 116-117).

beddi (it. belli ; litt. beaux), dans lequel nous avons constaté l'altération de l'ancien groupe consonantique -LL- en *dd-* telle qu'elle a été observée dans les parlers siciliens (Devoto *et al.*, 2002 : 146).

Cependant, outre l'emploi de l'adjectif *beddi*, nous avons observé l'usage dans le même texte, à deux reprises, de la forme *bella* (it. bella ; litt. belle), ce qui suppose un emploi alterné de la forme dialectale et de la forme italianisée. Nous avons remarqué ce genre de variation de manière répétée dans d'autres textes.

Une réponse plausible à ce problème serait que nous sommes en présence d'un phénomène de variation graphique, spécifique des écritures dialectales (Jejcic, 2006). Il se pourrait aussi que ce fait soit révélateur d'une influence de plus en plus importante de la langue italienne au début du XXe siècle, due notamment à la montée en puissance du fascisme, et, par conséquent, à un amoindrissement de l'emploi des dialectes et parlers.

L'étude de ce corpus multilingue est également destinée à mettre en lumière les phénomènes de contacts entre diverses variétés de langues et à examiner plus particulièrement le fonctionnement de l'emprunt à l'arabe tunisien dans le tissu syntaxique des chroniques (Mancini, 1994 ; Pellegrini, 1972). Nous souhaitons analyser les critères d'intégration, le degré d'assimilation et les modalités d'insertion des mots arabo-tunisiens (Deroy, 1980).

A ce sujet, nous avons observé l'adoption de certains emprunts attendus tels que des référents à des objets quotidiens. Nous avons l'exemple de *hari*[12], emprunt au substantif arabe *hāra,* employé dans le dialecte tunisien pour désigner une quantité précise d'œufs, soit « quatre œufs ».

Ce terme n'a pas d'équivalent en dialecte sicilien, ni en langue italienne, et a été assimilé sans changer de catégorie grammaticale dans la phrase repérée dans le document. Ainsi, nous avons l'emploi suivant :

(1) *Quattru* ***hari*** *e menzu comu l'ova* (1924_658_2_S.)

It. Quattro < quattro uova > e mezzo come le uova

Litt. Quatre < quatre œufs > et demi comme les oeufs

Nous avons également repéré des insertions pragmatiques tels que des adverbes, des verbes, ainsi que des expressions idiomatiques typiquement tunisiennes tels que *Mabruccu*[13] (it. auguri ; litt. félicitations), *Scialla*[14] (it. speriamo ! se Dio vuole ; litt. espérons !, si Dieu veut !).

Ainsi, ce type d'emprunt linguistique témoigne du degré d'intégration de la communauté italienne, plus particulièrement des Siciliens, dans le tissu social arabo-tunisien malgré les différences de religion, de culture et de langue.

Nous constatons que les chroniques composant le corpus se caractérisent par la diversité de leurs signataires. Une analyse approfondie des spécificités de chaque texte pourrait permettre, par exemple, de mettre en lumière la présence de plusieurs idiolectes.

Notre travail de recherche pourrait aussi nous fournir de précieuses indications sur la nature des rapports sociaux entre les diverses communautés vivant dans la Tunisie coloniale, et nous permettre de prouver que des facteurs extralinguistiques (cohabitation dans les mêmes quartiers, appartenance à des classes sociales similaires, etc.) ont eu un impact significatif sur les échanges linguistiques entre la variété dialectale d'arabe tunisien et le parler des Siciliens de Tunisie.

Dans la partie qui suit, nous exposons les outils et méthodes que nous souhaitons utiliser dans le traitement de notre corpus, ainsi que les problèmes rencontrés.

---

[12] « Doppu lu futtiballi (*Sceni di lu veru*) ». In *Simpaticuni*, 31 mai 1924, n°658, p. 2 (signé *Schut*).

[13] « Il Boicottaggio a Tunisi (*Sceni di lu veru*) ». In *Simpaticuni*, 28 et 29 octobre 1911, n°8, pp.1-2 (signé *R. C.*).

[14] « Il Boicottaggio a Tunisi (*Sceni di lu veru*) ». In *Simpaticuni*, 28 et 29 octobre 1911, n°8, pp.1-2 (signé *R. C.*).

4 Outils, méthodes et problèmes de traitement du corpus

4.1 Du grand format à la saisie manuelle

Le premier traitement nécessaire est la numérisation du corpus. Cette première étape nous semble incontournable pour pouvoir mener à bien nos observations.

Ainsi, seule la numérisation puis le traitement informatique du corpus pourraient aboutir au repérage de formes spécifiques, et à une analyse originale et exhaustive.

D'un point de vue technique, la numérisation pose toutefois certaines difficultés. En effet, nous sommes confrontés au problème de la dimension des feuilles du journal dont le grand format[15] complique la saisie automatique. Une solution possible serait de réaliser plusieurs prises de vue, soit d'effectuer un traitement par zones de textes, afin d'obtenir la meilleure retranscription possible des documents papier en format *TIFF* (*Tagged Image File Format*). Néanmoins, ce travail pourrait prendre un temps considérable et ne pas aboutir à un résultat satisfaisant.

Un autre problème technique auquel nous sommes confrontés est la publication, dans le sens vertical, des textes, et dans certains cas, la division d'un même texte sur plusieurs colonnes.

Une solution possible serait de réaliser la numérisation, puis d'utiliser un logiciel approprié permettant d'effectuer du découpage informatique afin de pouvoir reconstituer des fichiers correspondant aux textes scannés.

Outre la taille des feuilles, l'existence, dans le même texte, de traits dialectaux et de variétés de langues différentes, non normées pour certaines, pose le problème de la reconnaissance des caractères et, par conséquent, de la restitution fidèle du document original par un logiciel. Un OCR est donc inopérant.

La saisie manuelle au clavier a donc été la seule solution. Cette méthode a du moins l'avantage de résoudre les problèmes de numérisation et de reconnaissance des caractères, et permet ainsi une retranscription fidèle des textes originaux.

Notre corpus compte actuellement 220 textes, soit 162. 000 mots approximativement.

Lors de l'établissement de la version numérique du corpus, nous n'avons procédé à aucune normalisation orthographique et avons scrupuleusement respecté l'orthographe originelle.

Pour le moment, les enrichissements textuels (italique, gras, etc.) et les majuscules ont été sauvegardés dans le but de retranscrire fidèlement les documents imprimés. Au cours du travail d'analyse du corpus, nous déciderons s'il est nécessaire de sauvegarder ces caractéristiques, ou s'il est plutôt préférable de les négliger.

4.2 Représentativité du corpus

Un autre problème auquel nous sommes confrontés est celui de l'étendue du corpus et, par conséquent, de sa représentativité, partielle ou dans sa totalité. En effet, composé de 162. 000 mots approximativement, il est volumineux.

Est-il plus judicieux de procéder à la sélection d'un groupe de textes afin de constituer un échantillon qui serait représentatif de l'ensemble du corpus ? Ou bien doit-on embrasser la totalité et procéder ainsi à l'analyse de tous les documents recueillis ?

Dans certains cas, le choix d'un échantillon représentatif pourrait convenir quand les observations tendent à se recouper, en d'autres termes quand les résultats obtenus peuvent être projetés sur

[15] De l'année 1911 jusqu'à l'année 1913, le journal avait opté pour le format A3 (format *Tabloïd*). A partir de 1914 jusqu'en 1928, le journal a adopté un format plus grand dont les dimensions, 522 × 379 mm environ, se situent entre celles du format *Belge* (520 × 365 mm) et celles du Grand Format (575 × 410 mm). Enfin, un nouveau format, dont les dimensions sont de 452 ×359 mm, a été choisi pour l'édition des numéros de 1932 et 1933.

l'ensemble des données considérées. L'essentiel est que cet échantillon atteigne une certaine validité scientifique sur une base quantitative minimale (Guiraud, 1960).

Cependant, dans un souci d'exhaustivité, nous pensons qu'une analyse statistique, englobant l'ensemble des textes composant le corpus, pourrait aboutir à une interprétation objective des phénomènes de variations linguistiques observés dans la langue hybride des chroniques du journal.

4.3 Outils envisagés

Sur un plan méthodologique, nous envisageons d'employer certains logiciels. Nous avons adopté le format Txt (texte brut et non structuré) afin d'obtenir une compatibilité avec les outils choisis.

4.3.1 Lexico3

L'un des logiciels que nous souhaitons utiliser est *Lexico3.* Une des fonctionnalités de ce logiciel d'analyse des données textuelles est de pouvoir traiter diverses parties d'un même texte.

Il serait ainsi intéressant de vérifier les usages linguistiques dans les différents documents analysés et, plus particulièrement, la variation (Jejcic, 1996).

En effet, l'un des problèmes techniques auquel nous sommes confrontés est l'organisation des textes composant le corpus en dialogues et la présence dans chaque texte de deux personnages, voire plus dans certains cas.

Ainsi, *Lexico3* pourrait nous donner des indications sur les points de similitude et de différence afin de visualiser les principales oppositions dans un texte donné (Habert, 2005).

Le découpage des différents textes en formes graphiques permettra d'extraire les formes spécifiques des diverses chroniques et de repérer les segments répétés.

Etant donné que les auteurs de cette chronique diffèrent d'un numéro à un autre, nous voudrions observer la fréquence d'usage des emprunts à l'arabe tunisien par exemple, ainsi que les distinctions pouvant exister entre les formes extraites.

A titre d'exemple, nous avons remarqué que l'auteur *R. C.* (initiales de *Rosario Cunsolo*) tend à utiliser des mots empruntés au parler tunisien dans ses chroniques, rédigées essentiellement en 1911 et en 1912. Cependant, *Mastru Mbrogghia*, auteur habitué des colonnes du journal *Simpaticuni* et ayant rédigé un nombre considérable de chroniques, n'utilise pratiquement jamais de termes provenant du tunisien. A ce stade, seule une analyse statistique exhaustive pourrait aboutir à des résultats satisfaisants permettant une interprétation objective des faits

Ainsi, nous souhaitons que cette méthode nous permette d'établir des comparaisons afin de vérifier si l'emploi de divers idiolectes dans la rédaction de cette chronique est avéré.

Le logiciel *Lexico3* propose également l'observation des occurrences d'une forme ou d'une famille de formes en contexte. Cette fonction pourrait permette l'observation de la variation des occurrences dans son milieu. Nous pourrions également vérifier l'existence de corrélations entre la forme d'une occurrence cible et les formes de son environnement immédiat (Jejcic, 1996). A titre d'exemple, il serait intéressant d'observer l'emploi des emprunts au dialecte tunisien.

4.3.2 Unitex

L'emploi du logiciel *Unitex* est envisagé dans le but d'établir des concordances qui permettraient d'observer, de façon systématique et rigoureuse, l'insertion des mots et phrasèmes.

Nous avons remarqué l'usage, dans la langue hybride des textes composant le corpus, d'expressions idiomatiques ou figées, qui pourraient constituer des claques d'expressions arabes ou françaises. Il s'agit pour l'instant d'une hypothèse, d'où l'intérêt d'utiliser *Unitex* afin de vérifier les contextes d'emploi de ces expressions particulières.

Nous avons trouvé des expressions figées dont la structure aurait été calquée ou empruntée à d'autres langues. Voici un exemple repéré dans un texte de l'auteur *Marco Visconti* :

(2) *[...] Si scippanu li capiddi [...]* (1933_1080_1_M.V.)

It. […] Si mettono le mani nei capelli […], ou bien, Si strappano i capelli […]

Litt. […] Ils s'arrachent les cheveux […]

Nous avons mentionné plus haut le fait que nous avions repéré l'utilisation d'expressions idiomatiques d'origine arabo-tunisienne dans nos documents tels que l'interjection *Ia hasra*[16]*.* La forme arabo-tunisienne *yā hasra* permet au locuteur d'exprimer un sentiment de nostalgie, d'où son sens littéral *quelle époque, quelle nostalgie, que de souvenirs.* Ces expressions idiomatiques tunisiennes, qui ne possèdent pas d'équivalent dans la langue emprunteuse, sont couramment employées dans les textes et représentent ainsi un intérêt linguistique.

On constate également la présence d'expressions figées d'origine sicilienne ou méridionale comme dans l'exemple suivant :

(3) ***Orvu di l'occhi****, caru Gianni [...]* (1924_658_2_S.)

It. Cieco degli occhi, caro Gianni […]

Litt. Aveugle des yeux, cher Gianni […]

Cette expression possède un sens péjoratif. Elle est attestée en Sicile dans les dialectes de la province de Trapani (Piccitto, 1990, Vol. III : 413).

Cette expression, dont la forme varie sensiblement d'un texte à un autre, est fréquemment employée. Il serait intéressant d'en étudier le fonctionnement dans la syntaxe de la langue particulière des chroniques.

Par la suite, l'utilisation d'autres outils de traitement des corpus sera probablement nécessaire. Il serait intéressant par exemple d'employer d'autres instruments qui viendraient compléter notre analyse.

Est-il possible d'avoir recours à un étiqueteur morphosyntaxique dans le cas d'un corpus présentant d'importantes variations linguistiques ?

Existe-il d'autres analyseurs susceptibles de réaliser de nouveaux traitements statistiques sur nos textes ?

Nous souhaitons que ce colloque soit l'occasion d'échanger des idées et des expériences avec d'autres chercheurs, et qu'il puisse nous apporter des suggestions intéressantes et des réponses à nos questions.

4.4 Système de codage des textes du corpus

Les choix méthodologiques adoptés dans notre travail de recherche posent également le problème du codage des fichiers qui seront traités en continu pour plus de rapidité.

Nous voulons cependant conserver l'indication du fichier d'origine afin de retrouver l'emplacement des occurrences repérées, des segments isolés par les concordances, etc.

Il est par conséquent nécessaire d'adopter un système de nommage efficace qui doit permettre de revenir à tout moment au document original, et qui facilitera ainsi le travail de repérage lors de l'analyse du corpus par des outils informatiques (Rastier, 2011 : 210).

Dans le cas de notre corpus, l'en-tête de chaque texte correspond à un codage qui fait apparaître certaines données factuelles telles que l'année de parution, le numéro du journal et la ou les pages. Les initiales de chaque auteur ont été ajoutées afin d'éviter toute confusion en cas de présence de deux textes dans le même numéro.

[16]« Doppu lu futtiballi (*Sceni di lu veru*) ». In *Simpaticuni*, 31 mai 1924, n°658, p.2 (signé *Schut*).

A titre d'exemple, le codage figurant en en-tête du texte intitulé «Alla musica », publié en 1911 (9 juillet, n°2, p.2), et rédigé par *R. C.* (initiales de *Rosario Cunsolo*), est le suivant :

1911_2_2_R.C.

Eventuellement, des données complémentaires, relatives à la ligne et la colonne, pourraient être intégrées dans le système de nommage des fichiers par la suite.

5 Conclusion

L'apport de l'outil informatique et des logiciels contribue de façon non négligeable aux traitements des corpus.

Il n'en reste pas moins vrai que leur choix, leur mise en place et leur hiérarchisation dans le traitement des données s'avèrent difficiles.

Leur résolution est également cruciale puisqu'ils fondent la pertinence des analyses. En effet, l'emploi de ces instruments appelle une réflexion théorique plus poussée, ce qui pourrait être intéressant et tout à fait pertinent dans le cadre de notre recherche.

Fondé sur la complémentarité fonctionnelle des logiciels, notre travail de thèse pourrait aboutir au croisement des résultats de plusieurs méthodes, et de faire apparaître ainsi de nouveaux observables (Valette, 2008).

ALFIERI, G. (1992). La Sicilia. *In* (Bruni, F., 1992). *L'italiano nelle regioni. Lingua nazionale e identità regionali*, Torino. Unione Tipografico-Editrice Torinese, pages 798-860.

BARTENS, A. (2000). Vers une typologie socio- et psycholinguistique des produits du contact linguistique: exemples romans. *In* (Englebert, A., Pierrard, M., Rosier, L. et Van Raemdonck, D., éditeurs, 2000). *ACILPR XXII. Vol. 9: Contacts interlinguistiques*, Tübingen. Niemeyer, pages 7-18.

BERRUTO, G. (2002). *Sociolinguistica dell'italiano contemporaneo*, Roma. Carocci.

BILGER, M., éditeur (2000). *Corpus. Méthodologie et applications linguistiques*, Paris. Honoré Champion / Presses Universitaires de Perpignan.

BRONDINO, M. (1998). *La stampa italiana in Tunisia. Storia e società (1838-1956)*, Milano. Jaca Book.

COHEN, D. (1964). *Le parler arabe des Juifs de Tunis. Textes et documents linguistiques et ethnographiques,* Paris / La Haye. Mouton, Tome I.

DAVI, L. (2000). Entre colonisateurs et colonisés: les Italiens de Tunisie (XIX^{e}-XX^{e} siècles). *In* (Alexandropoulos, J. et Cabanel, P., éditeurs 2000). *La Tunisie mosaïque: Diasporas, cosmopolitisme, archéologies de l'identité*, Toulouse. Presses Universitaires du Mirail, pages 99-113.

DE MAURO, T. (2005). *Storia linguistica dell'Italia unita,* Roma / Bari. Laterza.

DEROY, L. (1980). *L'emprunt linguistique*, Paris. Les Belles Lettres.

DEVOTO, G. et GIACOMELLI, G., éditeurs (2002). *I dialetti delle regioni d'Italia,* Milano. Tascabili Bompiani.

GRASSI, C., SOBRERO, A. A. et TELMON, T., éditeurs (2005). *Fondamenti di dialettologia italiana*, Roma/Bari. Laterza.

GUIRAUD, P. (1960). *Problèmes et méthodes de la statistique linguistique,* Paris. Presses Universitaires de France.

HABERT, B. (2005). *Instruments et ressources électroniques pour le français*, Paris. Ophrys.

JEJCIC, F. (1996). L'écriture de variétés de français d'oïl : approche plurielle pour le traitement informatisé des variantes graphiques de textes patois. *In* (Moracchini, G., 1996). *Bases de données linguistiques : conceptions, réalisations, exploitations. Actes du colloque international de Corte (11-14 octobre 1995)*, Corte. Université de Corse, pages 277-293.

JEJCIC, F. (2006). Images de variétés de français du domaine d'oïl central : dynamiques de représentations graphiques d'auteurs (1911-1997). *In* (Jagueneau, L. éditeur, 2006). *Images et dynamiques de la langue. Poitevin-saintongeais, français et autres langues en situation de contact. Actes du colloque (Poitou-Charentes-Vendée, 6 décembre 2003 ; Poitiers, 5-6 novembre 2004)*, Paris. *L'Harmattan, pages 219-256.*

KOULOUGHLI, D. -E. (2007). *L'arabe*, Paris. Presses Universitaires de France.

LAKHDHAR, A. (2006). Fenomeni di contatto linguistico in Tunisia : la parlata mista dei siciliani di Tunisi e gli italianismi nella varietà dialettale di arabo tunisino. *In* (Banfi, E. et Iannàccaro, G. éditeurs, 2006). *Lo spazio linguistico italiano e le « lingue esotiche ». Rapporti e reciproci influssi. Atti del XXXIX Congresso di Studi della SLI (Milano, 22-24 settembre 2005)*, Roma. Bulzoni, pages 371-394.

LOTH, G. (1905). *Le peuplement italien en Tunisie et en Algérie*, Paris. Armand Colin.

MANCINI, M. (1994). Voci orientali ed esotiche nella lingua italiana. *In* (Serianni, L. et Trifone P. éditeurs, 1994). *Storia della lingua italiana. Le altre lingue. Vol.III*, Torino. Einaudi, pages 825-879.

MOIRAND, S. (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre*, Paris. Presses Universitaires de France.

PELLEGRINI, G. B. (1972). *Gli arabismi nelle lingue neolatine con speciale riguardo all'Italia*, Brescia. Paideia, 2 Volumi.

PENDOLA, M. (2000a). La lingua degli Italiani di Tunisia. *In* (Finzi, S., éditeur, 2000a). *Memorie italiane di Tunisia/Mémoires italiennes de Tunisie*, Tunis. Finzi, pages 13-18.

PENDOLA, M. (2000b). Les mots de la mémoire. Une approche linguistique de la présence italienne en Tunisie. *In* (Adda, L. éditeur, 2000b). *Les Relations tuniso-italiennes dans le contexte du protectorat. Actes du colloque international (Tunis, 12-13 mars 1999)*, Tunis. Institut Supérieur d'Histoire du Mouvement National/Université Tunis I, pages 83-93.

PENDOLA, M. (2000c). Mangiare all'italiana in Tunisia : un caso di sincretismo culturale. *In* (Finzi, S., éditeur, 2000a). *Memorie italiane di Tunisia/Mémoires italiennes de Tunisie*, Tunis. Finzi, pages 225-232.

PICCITTO, G., TROPEA, G. et SALVATORE, C., éditeurs (1977-2002). *Vocabolario Siciliano,* Catania/Palermo. Centro di Studi Filologici e Linguistici Siciliani, 5 Volumi.

RASTIER, F. (2011). *La mesure et le grain. Sémantique de corpus*, Paris. Honoré Champion.

SILBERZTEIN, M. (1998/1999). Traitement des expressions figées avec *INTEX*. *Linguisticae Investigationes,* Tome XXII, pages 425-449.

SOMAI, A. (2000a). Gli italiani di Tunisia attraverso la stampa umoristico-dialettale. L'esempio di « Simpaticuni ». *In* (Finzi, S., éditeur, 2000a). *Memorie italiane di Tunisia/Mémoires italiennes de Tunisie*, Tunis. Finzi, pages 189-210.

SOMAI, A. (2000b). Le laboratoire linguistique du journal SIMPATICUNI. *In* (Adda, L. éditeur, 2000b). *Les Relations tuniso-italiennes dans le contexte du protectorat. Actes du colloque international (Tunis, 12-13 mars 1999)*, Tunis. Institut Supérieur d'Histoire du Mouvement National/Université Tunis I, pages 193-209.

VALETTE, M., éditeur (2008). *Textes, documents numériques, corpus. Pour une science des textes instrumentée.* N°9, *Syntaxe et Sémantique*, Caen. Presses Universitaires de Caen.