Masarykova univerzita
Filozofická fakulta

Ústav románských jazyků a literatur

Entre son et texte : Analyse des versions parallèles des textes des chansons de rap avec des outils informatiques

Magisterská oborová práce

Vedoucí magisterské oborové práce:
PhDr. Alena Polická, Ph.D.

Autor:
Ing. Šárka Vaňková

Brno, 2014

Prohlašuji, že jsem magisterskou oborovou práci vypracovala samostatně s využitím uvedených pramenů a literatury a že se elektronická verze plně shoduje s verzí tištěnou.

V Brně dne

……

Šárka Vaňková

Je tiens à remercier à Madame Alena Polická, directrice de ce mémoire mineur, pour son soutien, sa disponibilité et ses remarques et conseils précieux qu'elle m'a apportés tout au long la rédaction de ce mémoire.

TABLE DES MATIÈRES

INTRODUCTION

La linguistique de corpus est le domaine de la linguistique informatique, un champ interdisciplinaire. Ses méthodes de recherche sont basées sur les outils électroniques et cette liaison rend possible l'examination des textes plus effective et plus rapide qu'auparavant. Un développement continuel de ces outils ne cesse d'améliorer les modes de travail dans le cadre de cette discipline nouvelle. En quelques clics de souris, les linguistes peuvent définir leur corpus de travail et trouver les informations souhaitées en quelques secondes. Cependant, ce ne sont pas seulement les corpus linguistiques qui facilitent le travail des linguistes ; il existe également d'autres outils comme des programmes d'alignement, des étiqueteurs, des analyseurs syntaxiques. En somme, la linguistique de corpus devient de plus en plus utile et elle se pratique dans des universités comme une discipline autonome, récente et pour cela discutée. L'accroissement de sa popularité est accompagné par l'augmentation des ouvrages scientifiques dont le sujet traite ses méthodes de travail et de constitution des bases textuelles. Ce mémoire mineur s'occupera de ce sujet tout en s'orientant vers l'alignement parallèle unilingue.

L'objectif sera la description et l'exploitation du programme *MkAlign*, programme d'alignement, sur un corpus textuel. Ce dernier sera composé des chansons de rap, interprétées par divers chanteurs ou groupes français. Le corpus sera basé sur les chansons étudiées dans le cadre du cours Langage des jeunes. L'orientation de ce cours explique aussi le choix de la musique rap qui reflète le mieux le langage des jeunes des banlieues. Nous avons décidé de traiter ce sujet car nous avons souhaité de continuer dans la recherche sur le champ des corpus linguistiques et de lier ainsi à notre mémoire de licence, rédigé quelques années avant. Néanmoins c'était avant tout le cours Langage des jeunes qui a suscité notre intérêt pour l'alignement parallèle et qui nous a permis de développer nos connaissances sur ce domaine de la linguistique.

Le présent mémoire sera divisé en trois parties. La première présentera la linguistique de corpus en général, plus particulièrement les corpus linguistiques et leurs types, leur constitution et leur exploitation. Elle introduira les notions de base de corpus. La deuxième partie, la partie théorique, sera consacrée au programme *MkAlign*, conçu par l'Université Sorbonne Nouvelle - Paris 3, à la description de ses fonctions et du procédé de l'enregistrement et du traitement des textes alignés. L'objectif en sera

de familiariser les futurs collaborateurs de la base RapCor avec le fonctionnement et les possibilités de ce programme. Un petit guide simplifié en tchèque sera une application pratique qui aidera toute personne intéressée à mener à bien son analyse. La personne en question sera ensuite mieux préparée à aborder l'analyse des textes choisis, la comparaison de la version « pochette » à la version « son » de la même chanson de rap. Cette même comparaison représentera le contenu de la troisième partie, à savoir la partie pratique de notre mémoire mineur.

Pendant la rédaction de la troisième partie, nous prêterons notre attention aux passages qui ne sont pas identiques et nous chercherons à déterminer la qualité des différences entre le texte écrit et le texte réellement chanté (p. ex. ce qui est omis, volontairement ou non, changé ou ajouté par le rappeur lui-même au cours de son interprétation). Nous examinerons ce sujet à partir de deux points de vue, conformément à deux hypothèses sur lesquelles le présent mémoire est basé. Premièrement, nous nous orienterons vers la structure des versions « texte » et « son », c'est-à-dire vers le côté formel, pour que nous puissions déterminer des parties omises ou, contrairement, ajoutées à la version « son » en comparaison à la version « pochette ». Nous supposons que le contenu des parties, étant ajoutées additionnellement au cours de l'interprétation de la chanson, est aussi impertinent, qu'il pourrait empêcher de publier une chanson touchée et pour cette raison, les chanteurs ne les mentionnent pas sur le livret de leur CD. Deuxièmement, nous traiterons ce sujet du côté textuel. Les chansons de rap peuvent être caractérisées comme un flux de paroles où l'enoncé représente l'élément le plus important. Nous présumons que les rappeurs ne sont pas trop soigneux et ils ne suivent pas le texte écrit sur le livret d'une manière stricte au cours de l'interprétation, même si leur enoncé reste le même. Nous tenterons de découvrir quelles sont les différences entre les versions « texte » et « son » sur le côté textuel et d'examiner si les rappeurs conservent l'énoncé, même s'ils font des changements au niveau des mots.

Comme le but de ce mémoire mineur est de déterminer et d'analyser les différences de façon quantitatif et qualitatif entre les versions « pochette » et « son » des diverses chansons de rap, l'échantillon analysé doit être suffisamment large. Le corpus examiné

sera donc composé des chansons de rap de la base de données du corpus RapCor[1] et il sera limité à cinquante chansons. Partiellement, il sera créé des chansons examinées pendant le cours Langage des jeunes (7 chansons). Le reste (43 chansons) sera complété par d'autres chansons sous format *.txt qui se trouvent dans le même centre de stockage. Pour cette deuxième source, nous prévoyons de choisir la sélection aléatoire comme méthode de travail.

Étant donné que ce mémoire mineur sera orienté sur le côté pratique, étant basé sur l'analyse des chansons, le texte sera accompagné de plusieurs graphiques et tableaux. Pour qu'on puisse juger la pertinence et l'exactitude de résultats présentés, nous ajouterons des documents les plus importants en annexes.

[1] Le corpus est disponible sur ce lien :
<https://is.muni.cz/auth/do/phil/Pracoviste/URJL/rapcor/library.html>.

1 LINGUISTIQUE DE CORPUS

Pour définir cette jeune discipline, prenons la définition formulée par František Čermák, directeur du Český národní korpus[2] et professeur de l'Université Charles à Prague : « La linguistique de corpus est la partie et la forme de la linguistique qui étudie la langue à l'aide des corpus et de la méthodologie qui en est déduite. »[3] Les corpus constituent donc la base de ce domaine linguistique. En général, il s'agit d'un ensemble d'exemples de l'utilisation du langage naturel, d'une certaine base matérielle servant à analyser et à décrire cette langue.[4] Dans le passé, les linguistes et leurs collaborateurs créaient ces corpus manuellement. Au cours des années, plus particulièrement depuis les années quatre-vingt, le procédé manuel a été remplacé par celui automatique. Ces textes, sous forme électronique, représentent une nouvelle base matérielle étant beaucoup plus large que l'ancienne. Les ordinateurs ont donc accéléré et facilité le travail des linguistes et le temps nécessaire pour traiter des documents s'est réduit au minimum. Ce processus n'est pas encore achevé, parce que les programmes électroniques, comme les lemmatiseurs, les concordanceurs ou les programmes d'alignement, ne cessent de se développer. Ce chapitre s'occupera d'abord du corpus en général étant la base de la linguistique de corpus, puis il abordera des programmes fondamentaux destinés au traitement de textes choisis.

1.1 Corpus

Aujourd'hui, on n'utilise le terme « corpus » en linguistique que pour le corpus électronique, soit « l'ensemble des textes stocké et conservé sous forme électronique »[5]. Même si les diverses définitions du corpus diffèrent en détail, elles partagent généralement trois points communs : le corpus devrait être a) d'un format homogène, b) fourni des tags linguistiques et c) d'un équilibre, doté donc d'une valeur de référence.[6] Il fait une partie intégrante de la linguistique appliquée dont le but est

[2] Corpus national tchèque (ČNK), projet de la Faculté des lettres de l'Université Charles, Prague.
[3] « Korpusová lingvistika je ta část a podoba lingvistiky, která studuje jazyk prostřednictvím korpusů a od nich odvozené metodologie. » František ČERMÁK, « Korpusová lingvistika dnešní doby ». In : František ČERMÁK et Renata BLATNÁ (éds), *Korpusová lingvistika: Stav a modelové přístupy*, Praha, Nakladatelství Lidové noviny, 2006, p. 9.
[4] Michal ŠULC, *Korpusová lingvistika: První vstup*, Praha, Nakladatelství Karolinum, 1999, p. 9.
[5] « (…) elektronicky uložený a uchovávaný soubor textů (…). » *Ibid.*, pp. 9-10.
[6] *Ibid.*, p. 11.

de décrire des langues naturelles et leurs variantes le mieux possible.[7] C'est aussi l'objectif des chercheurs sur le corpus. Des documents montrant l'utilisation authentique de la langue font le contenu du corpus et la recherche linguistique représente son but.[8] On distingue plusieurs types des corpus à buts différents.

1.1.1 Typologie des corpus

Étant donné que l'application des corpus est large et pluridisciplinaire, c'est-à-dire l'histoire, la sociologie, etc., il est évident qu'on en distingue plusieurs types. Dans ce chapitre, nous nous orientons vers la typologie des corpus en linguistique. La classification dépend d'un trait qu'on prend en considération ; cela peut être la taille d'un corpus, sa nature, sa chronologie, sa forme, sa ou ses langues. Cependant ces critères ne sont pas les seuls et ils peuvent se combiner mutuellement. La division fondamentale est celle linguistique, nous mentionnons également la division technique.[9]

La division linguistique tient compte plus particulièrement de la chronologie et de la forme des corpus. Sous le terme « chronologie » on comprend la synchronie et la diachronie. D'après les chercheurs[10] de l'Institut du ČNK, l'approche dite synchronique est plus courante dans les corpus et ce sont des corpus synchroniques et écrits qui prédominent aujourd'hui. Le point de vue de l'oralité ou de l'écriture, donc la forme des corpus, constitue également un des critères de la description (corpus oraux, corpus écrits).

Selon la taille, on différencie des grands corpus (l'étendue de centaines de millions de mots textuels), des corpus moyens (dix millions de mots textuels) et des petits corpus. En considérant la nature des corpus, Bowker et Pearson distinguent des corpus de référence et des corpus spécialisés. « Un corpus de référence reflète une langue et permet de faire des observations d'ordre général [...]. Un corpus spécialisé est axé sur l'aspect particulier du vocabulaire d'un domaine, sur un certain type de textes, sur

[7] Wolfgang TEUBERT, La linguistique de corpus : une alternative [version abrégée]. *Semen*, 27 [en ligne]. Le 2009-04-01 [page consultée le 2012-07-24]. Disponible à l'adresse : <http://semen.revues.org/8914>.
[8] M. ŠULC, *Korpusová lingvistika: První…*, *op. cit.*, p. 10.
[9] Jan KOCEK et Marie KOPŘIVOVÁ et Karel KUČERA (éds.), *Český národní korpus: Úvod a příručka uživatele,* Praha, ÚČNK FF UK, 2000, pp. 7 – 8.
[10] *Ibid.*, p. 7.

le langage des membres d'un groupe social [...]. »[11] Les deux types peuvent être soit ouverts, cela veut dire constamment élargissables, soit clos.

La question d'une langue ou des langues y représente un autre aspect jugé. On distingue des corpus monolingues et des corpus multilingues, mais cette division est basée sur des traits techniques plutôt que linguistiques. La classification technique différencie des corpus de contrôle, des corpus parallèles, des corpus apprenants et des corpus de test. Les corpus de contrôle sont d'habitude ouverts, ils prennent donc en considération l'évolution d'une langue. Les corpus parallèles sont au moins bilingues et ils contiennent toujours des originaux et leurs traductions. Ils sont exploités dans le but de la comparaison. Les corpus apprenants sont formés des textes provenant des étudiants de langues étrangères. Les corpus ne sont appliqués de cette manière-là que dans les derniers temps, mais il apparaît que l'analyse de ces corpus peut contribuer à l'enrichissement de l'enseignement à l'école. Les corpus d'entraînement et de test sont utilisés pour améliorer des programmes d'annotation et pour prouver des hypothèses linguistiques.[12]

1.1.2 Constitution d'un corpus électronique

Elizabeth Marshman[13] mentionne plusieurs facteurs à remplir pour qu'un corpus constitué soit de qualité et les résultats de la recherche soient adéquats. Ses exigences sont les suivantes :

- que le domaine des textes dans le corpus soit bien défini et délimité
- que les textes soient assez représentatifs pour appuyer les conclusions qu'on en tire
- que l'organisation, l'annotation, et le contenu du corpus favorisent son exploitation

Après avoir défini le but d'un corpus que nous projetons de construire, il faut choisir des textes y correspondant. Il faut les électroniser en les transcrivant par la méthode de l'OCR dans un format textuel. Il est nécessaire d'épurer ces textes (et ainsi éliminer des fautes créées liées à la transcription) et d'unifier leurs formats. Puis, les textes sont munis d'une annotation, de balises externes et internes si possible. Cependant,

[11] *1.2 Quel type de corpus constituer ?* [en ligne]. Page consultée le 2013-10-13. Disponible à l'adresse : <http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2005.ahronian_c&part=90677>.
[12] J. KOCEK et al (éds.), *Český národní korpus…*, *op. cit.*, pp. 8 – 9.
[13] Elizabeth MARSHMAN, *Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie* [en ligne]. 2003 [page consultée le 2013-10-13]. Disponible à l'adresse : <http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf>.

ce processus est trop exigeant en temps et en équipement technique pour un utilisateur courant et il semble beaucoup plus facile d'utiliser des corpus déjà existants - si cela est possible – qui sont déjà annotés. Elizabeth Marschman rappelle à propos de Bowker qu'il souligne :

> ...une importance d'un système de gestion polyvalent et complet afin de faciliter l'utilisation, le partage, et la modification de corpus. Il doit être possible d'adapter le corpus à des besoins de projets autres que celui pour lequel le corpus a été créé. Cette nouvelle application pourrait être similaire ou complètement différente de l'application originale.[14]

Le balisage externe permet de caractériser chaque texte d'une manière générale. Cette caractéristique contient les métadonnées, c'est-à-dire les informations telles que le nom de l'auteur, la date d'édition etc. Les logiciels traitant les corpus visualisent ces informations au bout de chaque ligne de concordance, sous forme d'un signe le plus souvent. Grâce à ces balises, des utilisateurs peuvent délimiter leur corpus de travail en utilisant un concordancier. Il est possible par exemple de trouver tous les textes créés dans une certaine année ou par un auteur concret.[15] Dans la plupart des corpus électroniques, ces informations additionnelles sont inscrites sous format SGML[16] ou XML[17] d'après une norme internationale TEI, soit « Text Encoding Initiative ».[18]

Le balisage externe ne doit pas être confondu avec le balisage interne. Ce dernier sert à définir la structure des textes et à apporter des informations linguistiques dans le corpus. Les balises externes comportent des informations sur la segmentation des textes aux chapitres, aux articles, aux phrases et aux mots. Le balisage interne contient le plus souvent un balisage morphologique des différentes formes de mot (des genres grammaticaux, y compris une lemmatisation) et ce balisage est inscrit dans le corpus à l'aide de programmes de segmentation et d'annotation automatique, par exemple TreeTagger. Toutes ces balises valorisent remarquablement le corpus.[19]

---

[14] Elizabeth MARSHMAN, *Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie* [en ligne]. 2003 [page consultée le 2013-10-13]. Disponible à l'adresse : <http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf>.
[15] J. KOCEK et al (éds.), *Český národní korpus…*, *op. cit.*, pp. 6 - 7.
[16] Standard Generalized Markup Language (language normalisé de balisage généralisé).
[17] Extensible Markup Language (langage de balisage extensible).
[18] F. ČERMÁK, « Korpusová lingvistika dnešní doby ». In : F. ČERMÁK et R. BLATNÁ (éds), *Korpusová lingvistika: Stav…*, art. cit., p. 12.
[19] J. KOCEK et al (éds.), *Český národní korpus…*, *op. cit.*, p. 7.

Quelques corpus comprennent également une décomposition analytique, « parsing » en anglais ; ces corpus sont munis de balises syntaxiques. En plus, les corpus parallèles sont traités par des programmes d'alignement dont le but est d'assurer l'alignement de parties de textes qui se correspondent mutuellement.[20] Les opérations mentionnées ne sont plus réalisées manuellement, mais elles se déroulent automatiquement ou, tout au moins, semi-automatiquement, grâce aux programmes qui seront décrits dans le sous-chapitre suivant. Avant de procéder à la description de ces programmes, une parenthèse sur l'exploitation des corpus nous semble nécessaire, en vue de notre partie pratique.

1.1.3 Exploitation des corpus

De grands corpus équilibrés servent de source d'informations aux spécialistes provenant de domaines variés ; les corpus peuvent être exploités par des littéraires, des sociologues, des psychologues, des historiens et bien sûr par des linguistes. Cependant, ce ne sont pas seulement les spécialistes qui travaillent avec les corpus ; les corpus sont destinés aussi aux étudiants.[21]

Le groupe le plus nombreux des utilisateurs de corpus est composé de linguistes. Ils utilisent des données de corpus pour plusieurs objectifs – pour la description de la langue, pour la création et la vérification de théories linguistiques et également pour la constitution d'applications basées sur ces données facilitant le travail aux lexicographes tout particulièrement. En comparaison avec les procédés précédents, celui-ci permet de consulter une information spécifique plus en détail et plus rapidement. En même temps, ce fait a mené à la croissance de la confiance en données déduites sur la base de corpus électroniques (les corpus d'aujourd'hui contiennent des centaines de millions de mots en comparaison des quinze millions de mots du fichier lexical, qui constituait la plus grande source auparavant).[22] On peut le démontrer à la concordance, soit une liste de toutes les lignes dans lesquelles le mot cherché se trouve. De cette manière, l'utilisateur voit le contexte du mot, limité par l'ordre de l'utilisateur. En plus, le programme de corpus permet de visualiser des caractéristiques de fréquence du mot. L'analyse de toutes ces données nous

[20] M. ŠULC, *Korpusová lingvistika: První…*, *op. cit.*, p. 13.
[21] J. KOCEK et al. (éds.), *Český národní korpus…*, *op. cit.*, p. 9.
[22] F. ČERMÁK, « Korpusová lingvistika dnešní doby ». In : F. ČERMÁK et R. BLATNÁ (éds), *Korpusová lingvistika: Stav…*, *art. cit.,* p. 15.

permettra de déterminer ce qui est typique et ce qui n'est que marginal dans la langue. Cette différenciation est importante avant tout pour les lexicographes dont le travail est justement de distinguer ces deux catégories. En outre, l'utilisateur peut comparer des résultats de plusieurs corpus et en déduire d'autres conclusions.[23]

1.2 Outils électroniques

Pour pouvoir constituer et exploiter les corpus, il faut disposer d'outils électroniques variés. Comme écrit ci-dessus, les corpus sont en grande partie construits à l'aide des programmes informatiques. Dans son article intitulé « Korpusy textů na FI MU », Pavel Rychlý[24] mentionne plusieurs manières d'enregistrer des corpus sur l'ordinateur. La forme la plus libre est représentée par des archives ou par une collection. Des textes faisant partie composante du corpus sont enregistrés sous format et codage variés, conformément à leurs documents source ; leur forme n'est donc pas unifiée. Une autre manière d'organiser les corpus est sous la forme d'une banque textuelle contenant des textes sous un même format et munis d'une annotation élémentaire (par exemple une détermination d'un type de source pour chaque article). La manière finale de l'enregistrement des corpus est représentée par l'utilisation d'un concordancier qui assure un codage des textes dans une certaine base de données. En résumé, les corpus comprennent les données suivantes : des textes, des métadonnées (concernant le nom d'un auteur des différents textes, leurs dates de publication, etc.), des informations sur la structure de chaque document (une différenciation des alinéas, des phrases, etc.) et une annotation (des informations sur les mots, la morphologie et les lemmes).[25]

Les textes dont le format est unifié, sont d'abord soumis à la tokenization ; ils sont décomposés en mots graphiques et en signes de ponctuation, c'est-à-dire aux tokens, les unités de segmentation. Ensuite, on procède à la segmentation, cela veut dire à la distinction des fins des phrases, à l'analyse morphologique et à la désambiguïsation pour éliminer la polysémie.[26] Les grands corpus disposent habituellement de leurs propres outils électroniques. En utilisant des corpus, les linguistes travaillent avec

[23] J. KOCEK et al (éds.), *Český národní korpus…*, *op. cit.*, pp. 9 – 10.
[24] Pavel RYCHLÝ, « Korpusy textů na FI MU », *Zpravodaj ÚVT MU*, 1997, année VIII, n° 2, pp. 9 - 12. ISSN 1212-0901.
[25] Karel PALA – Pavel RYCHLÝ, *Velké textové korpusy v praxi* [en ligne]. Le 2007-07-14 [page consultée le 2012-08-19]. Disponible à l'adresse : <http://www.datakon.cz/datakon08/d07_tut_pala.pdf>.
[26] Český národní korpus FF UK. *Korpus SYN* [en ligne]. Page consultée le 2012-08-20. Disponible à l'adresse : <http://ucnk.ff.cuni.cz/syn.php>.

un logiciel de recherche ou plutôt un concordancier[27] qui permet de rechercher un mot désiré. Toutes les occurrences de ce mot, mise en contexte, se visualisent sous format KWIC[28], le format le plus courant aujourd'hui pour la visualisation de lignes de concordance.[29]

En outre, il faut mentionner un lemmatiseur, « un programme de traitement automatique du langage qui permet de passer d'un mot portant des marques de flexion (pluriel, forme conjuguée d'un verbe...) à sa forme de référence (lemme ou forme canonique) ou inversement »[30]. Étant donné que chaque langue a ses propres particularités propres, il faut construire le lemmatiseur pour chaque langue particulière.

Un autre groupe de programmes nécessaires au traitement informatique de textes est celui d'étiqueteurs grammaticaux permettant de catégoriser des textes. Les textes, traités auparavant à la main, sont devenus un matériel d'entraînement pour ces programmes dont le but est de fournir des textes de corpus d'étiquettes de catégories grammaticales. Dans le cas de certains programmes, leur taux de réussite est de plus de 90 %.[31] Nous pouvons citer, comme exemple, *MorČe* et *Ajka* pour le tchèque et *TreeTagger* pour le français.

Les étiqueteurs grammaticaux sont souvent accompagnés de programmes servant à la désambiguïsation automatique des homographes. Leur fonction consiste à la sélection d'une catégorie grammaticale convenable selon le contexte d'un mot - homographe. De plus, on connaît des programmes à la décomposition syntaxique, capables d'analyser le texte et de l'étiqueter syntaxiquement (un parsage). Des corpus parallèles exploitent également des programmes d'alignement pour pouvoir joindre des parties (des phrases ou des alinéas le plus souvent) du texte original à des parties de la traduction y correspondant. L'un d'entre eux, *MkAlign*, fait l'objet du chapitre suivant.[32]

---

[27] F. ČERMÁK – J. KOCEK. *Co je korpus?* [en ligne]. Page consultée le 2012-08-24. Disponible à l'adresse : <http://ucnk.ff.cuni.cz/co_je_korpus.php>.
[28] Key Word in Context (Le mot clé dans le contexte)
[29] M . ŠULC, *Korpusová lingvistika: První…*, *op. cit.*, p. 20.
[30] Daniel GOUADEC, *Outils terminologiques* [en ligne]. Le 2010-10-20 [page consultée le 2012-08-23]. Disponible à l'adresse : <http://www.profession-traducteur.net/outils/outils.htm>.
[31] M. ŠULC, *Korpusová lingvistika: První…*, *op. cit.,* p. 21.
[32] *Ibid.*

2 PARTIE THÉORIQUE : MKALIGN

Le logiciel MkAlign est élaboré par le Centre de Lexicométrie et d'Analyse Automatique des Textes (SYLED-CLA2T) sous la responsabilité de Serge Fleury. Ce centre fait partie, avec le Centre de recherche sur les discours ordinaires et spécialisés (CEDISCOR) et le Centre de Recherche en Traductologie (CR-Trad), de l'équipe SYLED (Systèmes, Linguistiques, Énonciation et Discursivité) regroupant des enseignants-chercheurs. L'équipe SYLED se trouve à l'Université Sorbonne-Nouvelle, Paris 3.[33]

Selon la définition de Serge Fleury, « le programme MkAlign permet de construire, corriger et visualiser un alignement de deux textes via un éditeur à double entrée »[34], mais il ne s'agit pas seulement d'un aligneur automatique. Ce programme est conçu pour plusieurs buts et il permet de créer, d'aligner, de corriger et de valider des traductions, mais aussi de « mener des calculs lexicométriques sur les contenus textuels chargés »[35]. Un utilisateur non expérimenté de ce programme, un traducteur par exemple va apprécier la possibilité de visualiser l'alignement dans une représentation cartographique ou en anglais « bi-text map » affichant des similitudes au plan traductionnel.[36] Des autres informations concernant la description et l'utilisation du programme MkAlign sont accessibles à l'adresse web[37] où le manuel du programme se trouve.

2.1 Processus de fonctionnement et d'utilisation du programme MkAlign

Dans ce chapitre, nous traitons des fonctionnalités du programme *MkAlign* 2.00, la version 2.0b144-2, mise à jour le 5 juillet 2012. Nous puisons ces informations du manuel d'utilisation du *MkAlign*, la version 2.0 (b144), mise à jour en juin 2012. Étant donné que ce manuel a été préparé pour la version précédente du programme (le manuel actualisé n'est pas accessible au moment de la rédaction de ce mémoire

[33] *SYLED* [en ligne]. 2009 [page consultée le 2012-08-25]. Disponible à l'adresse : <http://syled.univ-paris3.fr/presentation.html>.
[34] Serge FLEURY, *MkAlign (version 2.0) : Manuel d'utilisation*. Paris, Université Sorbonne Nouvelle Paris 3, 2012, p. 10.
[35] *Ibid.*
[36] *Ibid.*
[37] Le lien est le suivant : <http://issuu.com/sfleury/docs/mkaligndoc/1>.

mineur), nous nous appuyons aussi sur la présentation du programme qui se trouve dans le volet ***home*** du programme. Après la présentation générale du programme, nous nous orienterons vers son utilisation pratique. Dans des sous-chapitres suivants, nous présenterons le processus de son exploitation sur des exemples concrets de notre corpus RapCor. Pour le démontrer, nous avons décidé d'enregistrer la chanson de rap « Salope.com » dont l'auteur est l'interprète Kool Shen. Cette chanson a été traduite lors du cours *Langage des jeunes : les problèmes de la sociolinguistique française*, mené par la directrice de notre mémoire, Mme Alena Polická.

« L'interface du programme est composée d'une fenêtre graphique disposant de différents onglets »[38], soit ***home, param, align, map, graphe, specif, coocs, liste, segment, concordance, variation, export-XML, export, export-L3, rapport***. Nous présenterons ici quelques-uns de ces onglets. L'onglet ***home*** apporte une présentation du programme en plusieurs points et une esquisse du procédé de travail avec des documents. L'onglet ***param*** « permet de modifier le paramétrage de certaines fonctionnalités du programme »[39], par exemple un signe du segmenteur, une taille des polices d'affichage, selon le besoin de l'utilisateur. D'autre part, il y en a quelques-unes qui sont invariables (par exemple le nombre de cellules alignées par page). L'onglet ***align*** est consacré au chargement des fichiers au programme. Ici, l'utilisateur décide entre trois modes d'alignement, un mode général, un mode d'alignement automatique par une recherche de cognats, ou un mode d'alignement au format TXM.[40] L'onglet ***map*** permet d'afficher une carte de l'alignement et de construire des calculs lexicométriques. L'onglet ***graphe*** fait apparaître une courbe d'accroissement du vocabulaire. L'onglet ***coocs*** offre la possibilité de calculer des cooccurrents et poly-cooccurrents. Le volet ***segments*** touche des segments répétés dans les textes examinés et le volet ***variation*** prend en considération « la mise à jour de la variation dans les textes »[41].[42]

Concernant les principales fonctionnalités de ce programme, on peut figer deux cellules alignées et empêcher ainsi de modifier leur contenu (la couleur des cellules est verte). Si une cellule est colorée en blanc, elle peut être modifiée en écriture. La couleur rose

[38] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 11.
[39] *Ibid.*, p. 12.
[40] *Ibid.*, p. 14.
[41] CLA^{2}T. MkAlign 2.00 (2.0b144-2) – l'onglet *home* (presentation partielle du programme).
[42] *Ibid.*

signale que la cellule est vide, créée après l'insertion d'un segmenteur.[43] L'insertion ou la suppression du segmenteur provoque un découpage ou une fusion automatique de la cellule concernée (et de la cellule suivante).[44] L'utilisateur lui-même peut choisir le caractère servant de segmenteur, il est possible de le paramétrer dans la section ***param***. S'il n'y a aucun caractère, c'est le caractère *retour chariot* qui prend fonction par défaut du segmenteur.[45] De plus, ce programme permet de lancer une recherche dans les textes alignés.

2.1.1 Paramètrage et enregistrement des textes au programme *MkAlign*

Comme mentionné ci-dessus, nous utilisons pour ce but la chanson « Salope.com » de l'interprète Kool Shen, plus particulièrement, deux textes, son original français (version son) et sa traduction en tchèque. Dans ce cas, les textes de travail seront chargés dans l'éditeur par un mode général. Ce mode exige de poursuivre trois démarches : de sélectionner un caractère du segmenteur, de charger premièrement un fichier *source* et deuxièmement un fichier *cible*.[46]

Il faut d'abord vérifier le paramétrage du programme dans l'onglet ***param***. Il y a au moins trois zones de saisie exigeant notre attention : le segmenteur, le codage du texte source et le codage du texte cible. Le segmenteur y est prédéfini, c'est le caractère #. Ce caractère est important, parce que les textes chargés sont ensuite divisés en segments sur cette base que nous conservons pour ce travail. Puis, il est nécessaire de choisir le codage des deux textes de travail. Dans les deux cas, nous avons choisi UTF-8 (Unicode)[47].

Une autre démarche consiste dans le traitement des textes de travail pour les préparer au chargement dans le programme *MkAlign*. Les textes doivent être nettoyés. S'il s'agit donc de textes de chanson, comme dans notre cas, il faut effacer les parties excessives pour l'objectif de ce programme (des mots comme « l'intro », « le couplet », etc.). Après, nous complétons les textes avec le caractère du segmenteur sélectionné.

[43] Voir la figure n°1.
[44] CLA^{2}T. MkAlign 2.00 (2.0b144-2) – l'onglet *home* (presentation partielle du programme).
[45] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 17.
[46] *Ibid.*, p. 14.
[47] Même s'il devrait être possible de choisir ISO 8859-1 (latin-1/West European) pour le français (texte source) et ISO 8859-2 (latin-2/Central European) pour le tchèque (texte cible), la visualisation de ces codages des caractères n'est pas sans problèmes et quelques caractères ne s'affichent pas correctement.

Sa répartition dépend de notre intention – si on souhaite découper les textes en phrases, il faut mettre le segmenteur après chaque phrase. Le schéma peut être le suivant[48] :

Fichier Source :
sssssssssssssssss #
sssssssssssssssss #
etc.

Fichier Cible :
ccccccccccccccc #
ccccccccccccccc #
etc.

C'est l'utilisateur lui-même qui définit la précision de l'alignement et ainsi la taille des segments. Dans notre travail, nous alignons au niveau des phrases. Les différentes phrases devraient succéder l'une après l'autre (sans lignes vides) pour que l'aligneur visualise les phrases alignées parallèlement. Si on affiche des caractères cachés (à l'aide de l'icône ¶ dans une barre d'outils), l'icône ¶ suit immédiatement le caractère # (les espaces éventuelles entre ces deux signes doivent être effacées). Le programme est capable de traiter des fichiers sous les formats *.xml, *.htm, *.html, *.txt. *.dic et *.par. Nos textes de travail sont sauvegardés en *.txt à l'aide du *Bloc-notes,* un éditeur de texte (le codage UTF-8). On peut ensuite accéder à l'alignement des textes de travail par l'intermédiaire du volet ***align***.

2.1.1.1 Mode général

Dans le volet ***align***, il faut activer le bouton  pour choisir et charger le texte source et le bouton  pour répéter cette operation avec le texte cible. Les deux textes sont visualisés et alignés automatiquement selon le caractère délimiteur. Si on souhaite de fixer le contenu de certaines cellules, on clique sur le bouton localisé à droite de chaque paire des cellules (la lettre W est remplacée par la lettre R et la couleur des cellules choisies est modifiée en vert)[49].

[48] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 15. « SSSSS » représent le texte source, ainsi que « ccccc » remplace le texte cible.
[49] Voir la troisième paire de cellules sur la figure n°1 : L'alignement des phrases et la fixation des cellules.

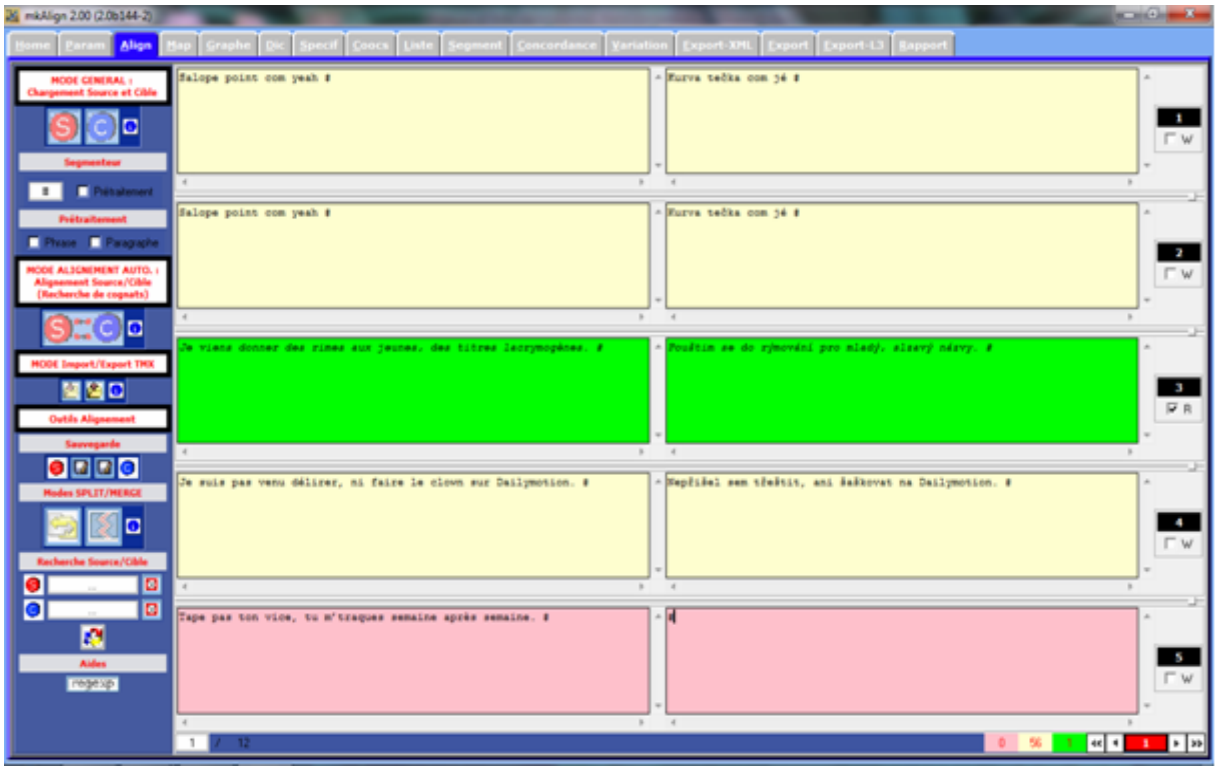


Figure n°1 : L'alignement des phrases et la fixation des cellules

Dans cette figure, il est possible de voir que « les 2 volets du corpus aligné sont présentés par page de 5 blocs alignés, on peut passer d'une page à l'autre de l'alignement via les boutons présents au bas de l'onglet ***align***, ou en sélectionnant une page donnée (puis touche Entrée) »[50]. Au bas de cet onglet, il y a des données numériques (le numéro de la page à laquelle l'utilisateur se trouve, le nombre de cellules vides, de cellules modifiables et de cellules protégées) qui aident l'utilisateur à s'orienter dans le programme.

Le mode général permet aussi de découper les textes de travail sans la nécessité d'insérer le segmenteur dans les textes de travail, il suffit d'activer la case à cocher *prétraitement*. Cette opération nous « permet de préformater les 2 textes à charger en phrases ou en paragraphes » sur la base des règles typographiques. Une phrase y est définie « comme une chaîne de caractères se terminant par les caractères suivants : point, 3 points, points d'interrogation et point d'exclamation », un paragraphe y est conçu « comme une suite de phrase terminé par un retour à la ligne ».[51] C'est à l'utilisateur de choisir lequel de ces deux types de pré-formatage il souhaitera activer. L'utilisateur peut ensuite préciser ou retravailler ce premier alignement.

[50] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 15.
[51] *Ibid.*, p. 18.

2.1.1.2 Mode « aligement par recherche de cognats »

Sous le terme de cognats, on conçoit « des mots qui partagent des propriétés phonologiques, orthographiques et sémantiques facilement repérables »[52], par ex. comprehension/compréhension pour l'anglais et le français.

Le mode « alignement par recherche de cognats » est un alignement automatique de fichiers qui est basé sur la recherche de points d'ancrage lexicaux. « Cette méthode permet, pour des langues apparentées, de construire un alignement en recherchant tout d'abord des équivalents traductionnels sous forme de mots apparentés (ou cognats) [...] »[53]. Avant de lancer ce type de l'alignement, il est demandé à l'utilisateur de remplir une fenêtre *Chargement des données pour l'alignement* affichée après avoir activé le bouton  . Il faut sélectionner ici la position des fichiers, leur encodage (le même pour tous les deux) et indiquer la langue des textes de travail. Puis, le processus d'alignement est lancé. Même si le français et le tchèque ne sont pas des langues trop apparentées, nos textes sont alignés presque correctement (surtout grâce aux mots intraduisibles, comme les noms propres[54]). À ce stade, il faut remarquer qu'en alignant des textes français et des textes tchèques, il serait plus convenable d'utiliser « le mode général ».

2.1.2 Outils de l'alignement

2.1.2.1. Modes

Pour modifier l'allure des cellules (le fractionnement et la fusion de la cellule)[55], on peut soit insérer/supprimer le caractère segmenteur, soit activer le bouton du mode *split*  ou le bouton du mode *merge*  (via un clic droit de la souris). Le clic gauche de la souris (à l'endroit où on veut réaliser le fractionnement ou la fusion) permet ensuite d'effectuer une opération souhaitée.[56]

[52] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 74.
[53] *Ibid.*, p. 22.
[54] Par ex. Dailymotion, Booska-p, Windows, YouTube.
[55] Le figement de cellule a été déjà traité ci-dessus.
[56] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 29.

2.1.2.2 Recherche de chaînes de caractères

La fonction de la recherche de chaînes de caractères permet de trouver des différents mots (ou une partie d'une phrase) ou des mots contenant une chaîne de caractères (définie par une expression régulière) dans les deux textes alignés. Si l'utilisateur veut élargir un champ de recherche, il insérera l'expression régulière dans la zone de saisie (puis appuiera sur ENTRÉE). La liste des opérateurs est disponible à la page 71[57] du manuel. Par exemple, l'instruction \bcou générera, dans l'original français de « Salope.com », des occurrences suivantes : couilles, coup, courant, cousin. Les occurrences sont colorées en rouge. Avant de lancer une nouvelle recherche, il est nécessaire d'effacer la précédente (par le bouton du rafraîchissement de l'éditeur ) ou de sauvegarder un sous-corpus construit des « cellules contenant le motif défini dans la zone de saisie »[58] (via le bouton ).[59]

2.1.3 Opérations réalisables grâce au programme *MkAlign*

L'onglet ***dic*** contient un dictionnaire des formes généré automatiquement après le chargement des textes de travail dans le programme. De plus, la liste est accompagnée de l'information sur la fréquence des différents mots ; elle est structurée, les mots sont rangés en ordre fréquentiel, des mots les plus fréquents aux mots n'ayant qu'une seule occurrence dans le texte.[60] Étant donné que les différents volets sont reliés entre eux, la recherche dans ce volet influence les résultats d'autres volets (l'onglet ***map***, ***graphe***...). Après avoir défini « un motif de recherche dans les zones de saisie *Recherche Forme(s)* »[61], l'utilisateur lance la recherche via la touche *enter*.[62]

---

[57] Cette page est disponible sur ce lien : <http://issuu.com/sfleury/docs/mkalignd oc/71>.
[58] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 27.
[59] *Ibid.*, pp. 26 – 27.
[60] Cette hiérarchie peut être changée par le clic-gauche de la souris sur le bouton *Fq* (soit la fréquence) ou sur le bouton *Forme* (les mots sont ensuite rangés dans l'ordre alphabétique).
[61] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 20.
[62] La même opération peut être lancée par le clic-gauche de la souris suivi du clic droit sur un mot choisi dans le dictionnaire (la colonne *Forme*), même si cette manière de la recherche est plus restreinte à cause de la nécessité de choisir un mot présent sur la liste (voir le côté droit de la figure n°2 – *Dictionnaire des formes (Cible)* – nous avons choisi le pronom personnel « se » et l'éditeur a trouvé toutes ses occurrences, y compris les mots contenant une chaîne de ces deux lettres comme « sebe, musej, etc. » ; pour spécifier notre instruction, il faut utiliser des expressions régulières (disponibles en anglais après le clic sur le bouton *regexp* qui réachemine l'utilisateur vers une page Web).

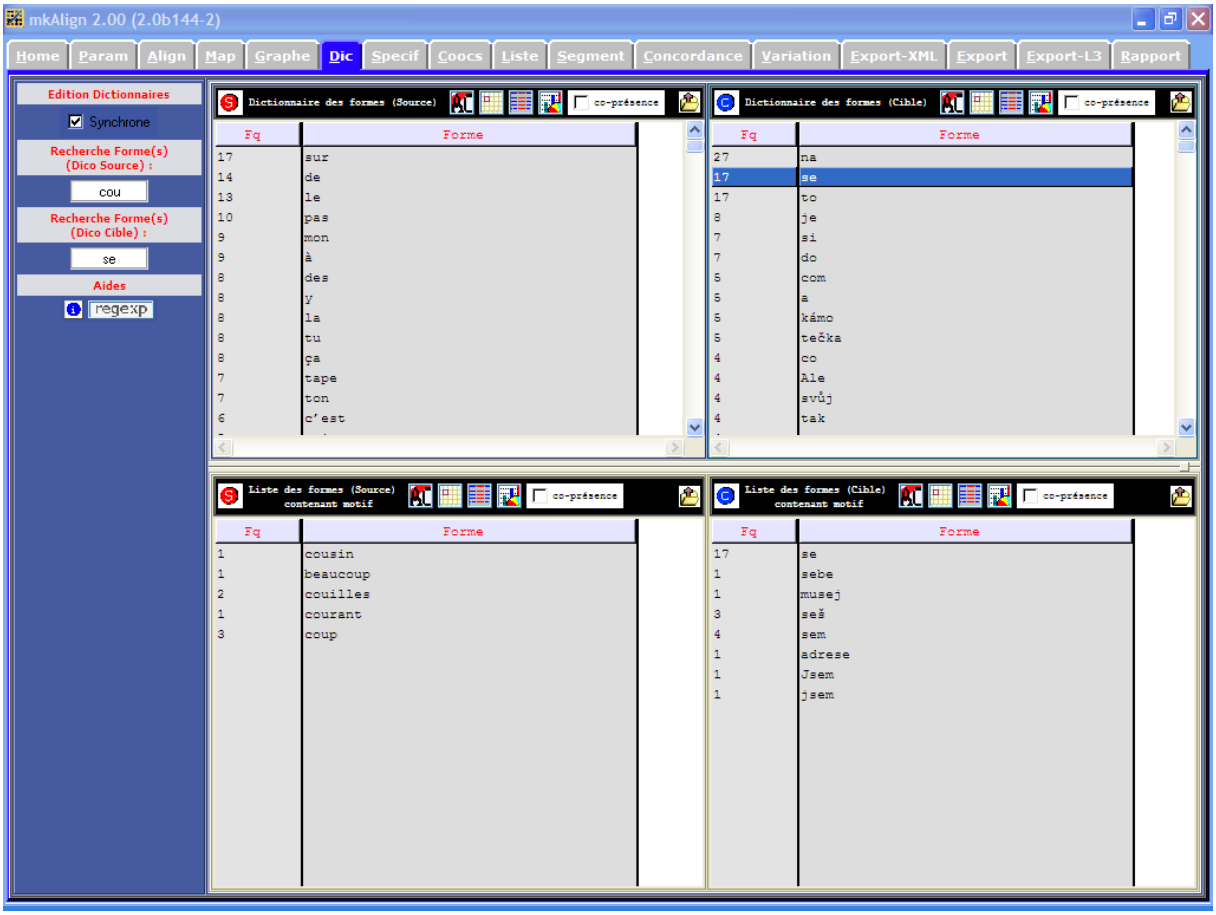


Figure n°2 : La recherche des occurrences dans les dictionnaires des formes (Source/Cible)

Chacune de quatre zones est munie d'une même série de boutons 

pour pouvoir être gérée individuellement. Le mot « coup » est selon la figure n°2 représenté trois fois dans le texte de travail français. Si l'utilisateur souhaite de connaître sa position exacte dans le texte, il le choisit via le clic gauche de la souris et puis, il clique sur l'icône  , située dans une barre d'outils mentionnée. Le programme le réachemine à l'onglet ***map*** où les résultats de la recherche sont affichés sous la forme d'une carte de l'alignement. Des carrés (ou des sections) sont toujours affiché(e)s par blocs de cinq. Chaque carré correspond à une cellule (soit cinq cellules alignées par page). Certains d'entre eux contiennent des diagonales indiquant le positionnement d'un mot recherché. Le résultat de la recherche se reflète aussi dans le volet cible représentant une zone miroir (le contour du carré aligné devient rouge)[63].

[63] La couleur d'un contour de carrés est différente si l'utilisateur sélectionne la recherche dans tous les deux volets, le texte source et le texte cible. Plusieurs informations sur les couleurs de carrés sont disponibles aux pages 32 et 33 du manuel.

Si l'utilisateur active une section via un clic gauche de la souris, le texte de cette section se visualise dans des zones d'édition au bas des cartes.[64]

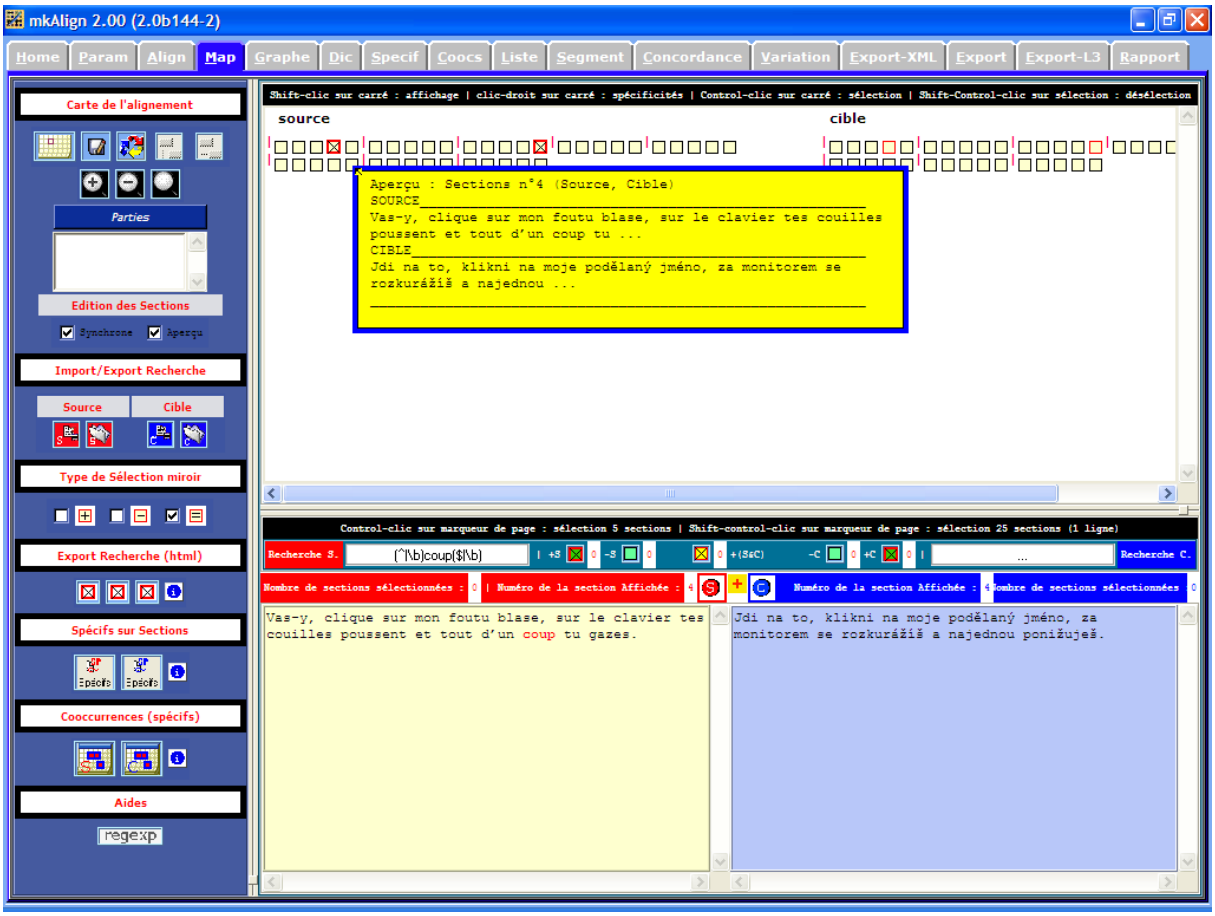


Figure n°3 : L'exemple de l'utilisation de l'onglet *map*

Sur la figure n°3, le texte blanc sur fond noir explique à l'utilisateur comment visualiser les particularités ou sélectionner les sections souhaitées. En suivant ces instructions, on active la combinaison de touches shift, control et clic gauche et on appuie sur le marqueur de page I, situé devant ces sections. Vingt-cinq sections, soit une ligne, sera sélectionnées. Puis, on activera le calcul du vocabulaire spécifique de cette sélection via les boutons *Spécifs sur sections* (soit Epécifs , soit Epécifs). Après cette opération, le programme nous réacheminera à l'onglet ***spécif*** où la figure n°4 s'affichera. « Le résultat produit donne à voir le vocabulaire spécifique des sections sélectionnées (dans la source ou dans la cible) et des sections associées (respectivement dans la cible ou dans la source) »[65].

[64] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 30.
[65] CLA^{2}T. MkAlign 2.00 (2.0b144-2) – l'onglet *map* (Aide spécifs).

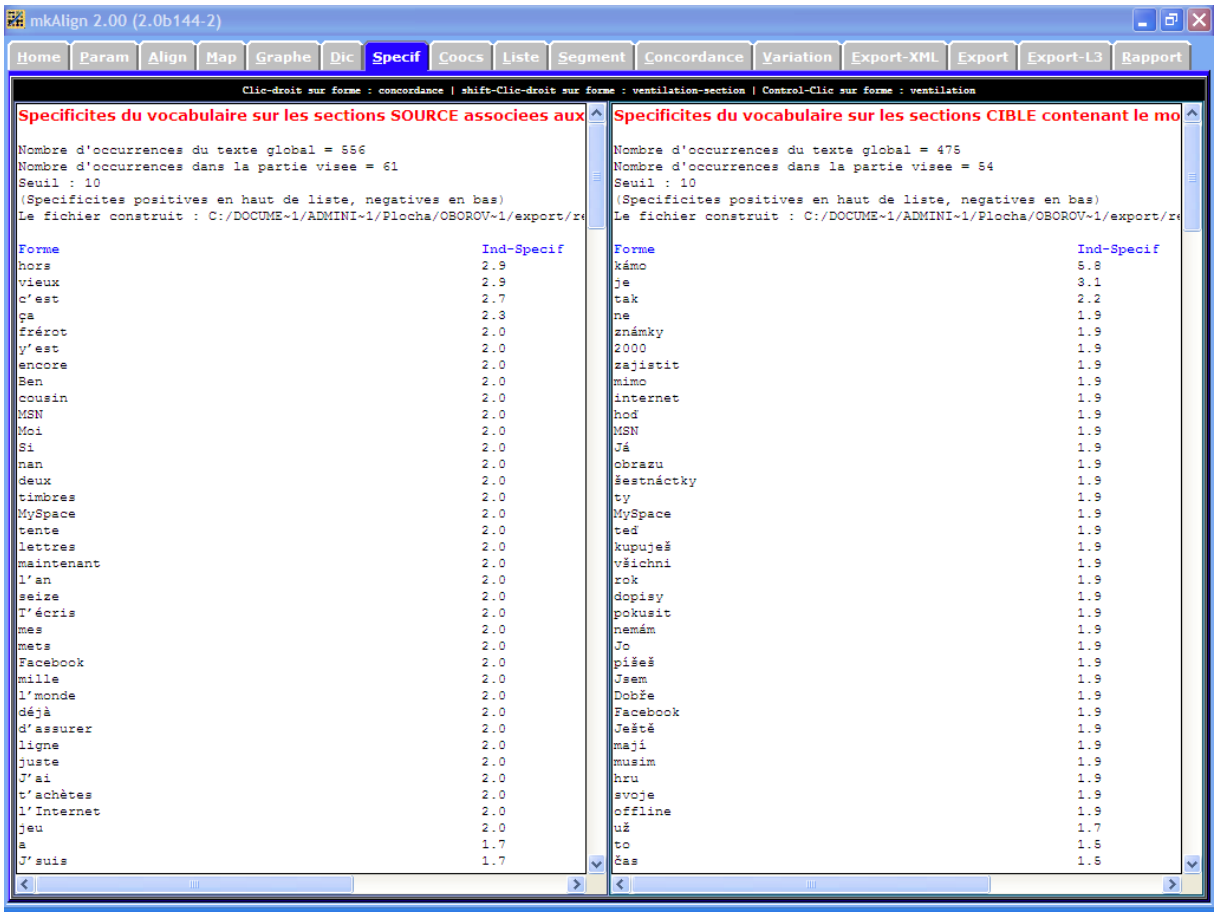


Figure n°4 : Visualisation du calcul du vocabulaire spécifique

Dans le volet ***dic***, le programme *MkAlign* offre la possibilité de définir des paramètres pour construire deux types de graphiques. D'abord, l'utilisateur fixe des mots dont l'occurrence devrait être examinée, puis il activera le bouton du graphique correspondant (*l'accroissement du vocabulaire sur forme sélectionnée* ou *la ventilation des formes sélectionnées*). Par exemple, nous avons décidé de comparer l'occurrence des mots *frérot*, *gros*, *mon vieux*, *frère*, *cousin* et leur traduction tchèque « kámo ». Toutes ces occurrences françaises n'ont pas été traduites d'un même équivalent tchèque, des courbes d'un graphique ne se correspondent pas donc dans tous les cas. Après une fixation des mots[66] cités ci-dessus et un clic gauche de la souris sur le bouton , le programme nous réacheminera à l'onglet ***graphe*** où on trouvera la figure suivante :

[66] Pour pouvoir fixer plusieurs mots dans le même texte (source/cible), il faut tenir la touche contrôle (Ctrl) appuyée en cliquant sur la souris.

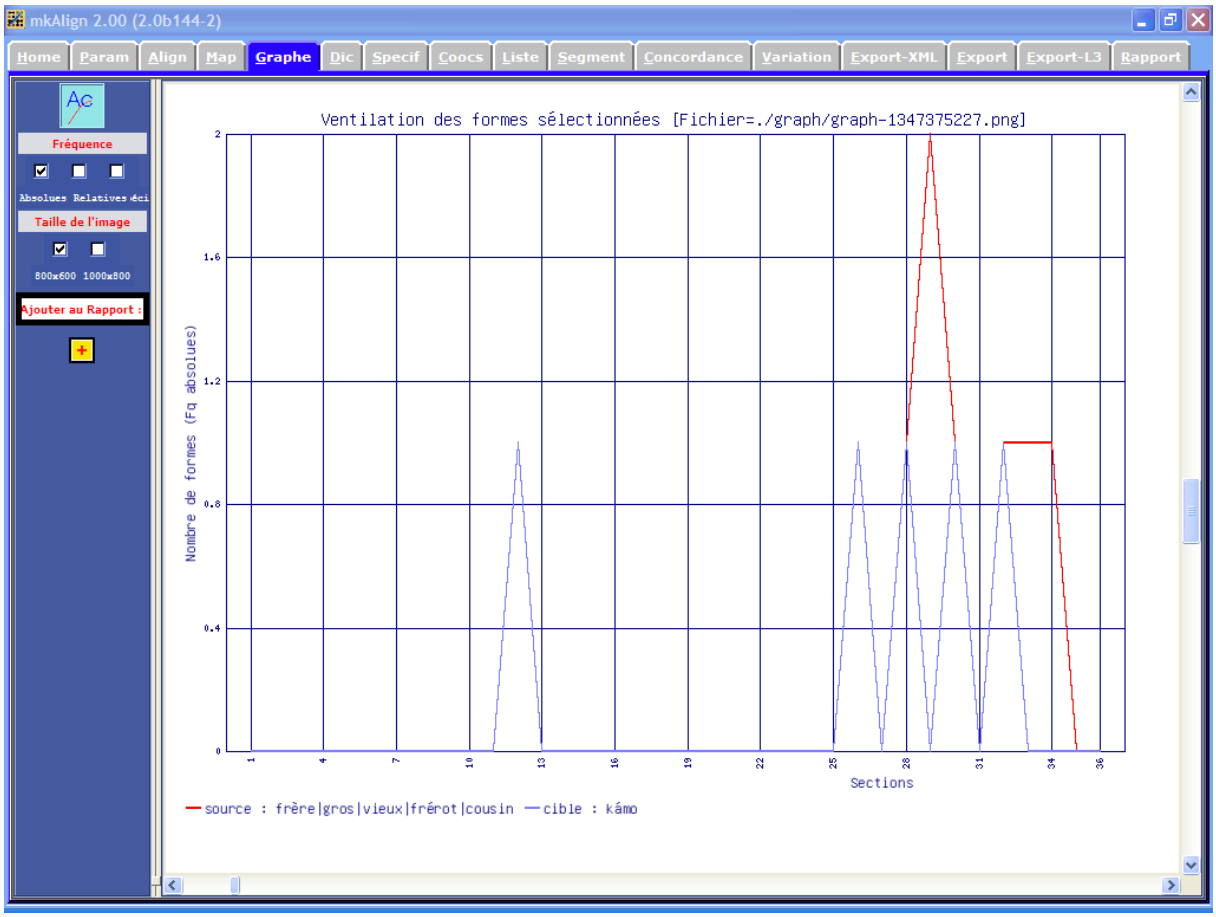


Figure n°4 : Graphique – la ventilation des formes sélectionnées

Un axe horizontal affiche les intervalles numériques de sections (le numéro de chaque troisième section), un axe vertical représente le nombre de formes (leurs fréquences absolues dans ce cas). Des courbes bleues et des courbes rouges se correspondent[67], sauf dans les cas des sections n°29, 33 et 34 où les mots *frère* (utilisé deux fois dans la section n°29), *gros* et de nouveau *frère* sont traduits d'une manière différente[68].

Après notre retour à l'onglet ***dic***, on peut modifier le paramétrage avant la construction du deuxième graphique. En le conservant, on est capable d'apprendre plusieurs informations sur les mots déjà recherchés. Nous avons décidé de conserver le paramétrage précédent en cliquant sur le bouton  . Le graphique de la courbe d'accroissement du vocabulaire est visualisé sur la figure n°5.

[67] La couleur finale est le bleu, même si ce fait n'est pas évident au premier aspect. Si l'occurrence n'était notée que dans le texte cible, une courbe bleue serait complétée d'une courbe rouge au niveau zéro sur un axe horizontal.
[68] Soit « brácho », « kamaráde ».

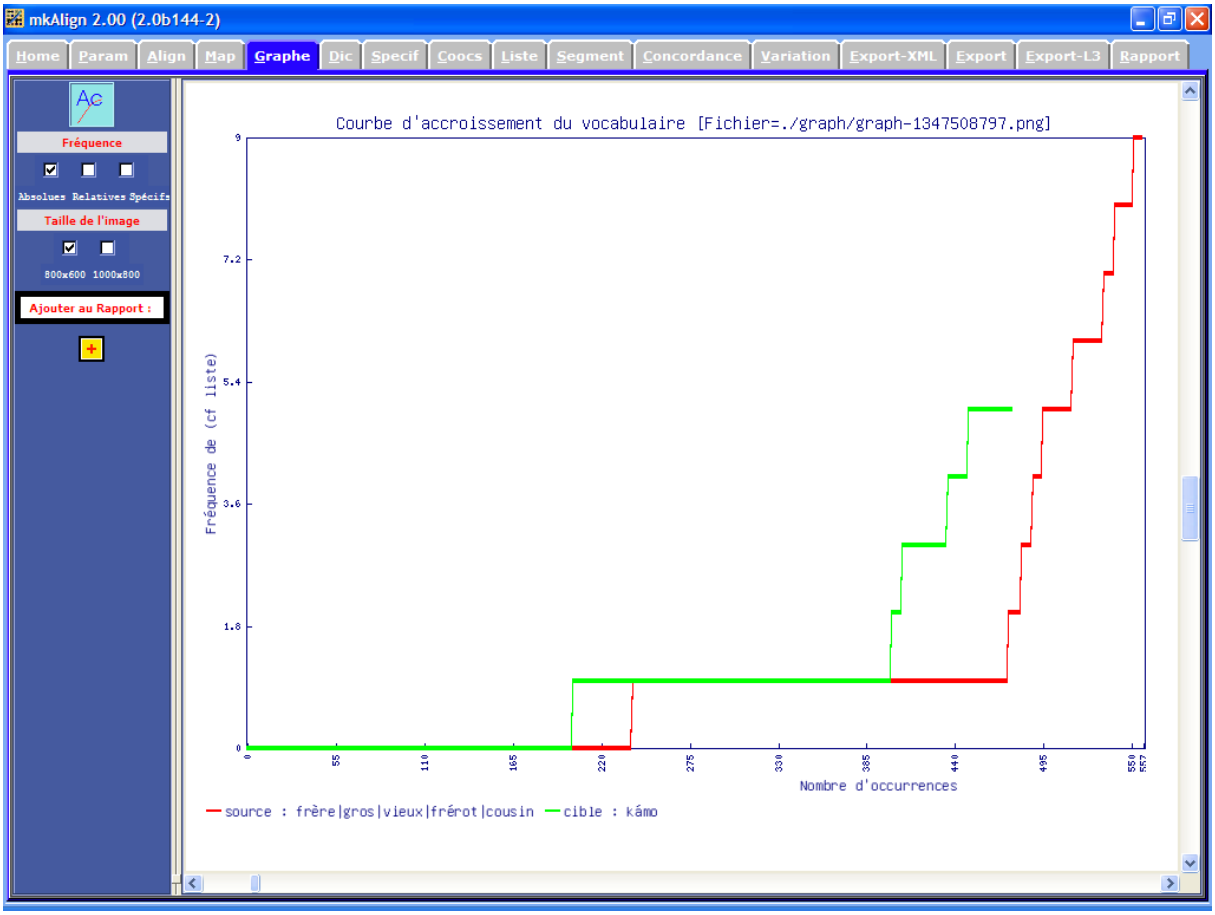


Figure n°5 : Graphique – le courbe d'accroissement du vocabulaire

Étant donné que le texte original de la chanson « Salope.com » est plus long que sa traduction tchèque, la courbe verte (le courbe cible) s'accroît plus tôt que la courbe rouge du texte source (un axe horizontal représente les différentes occurrences et non les différentes sections, comme dans le cas du graphique précédent). Sur la base du graphique, on peut constater que presque toutes les occurrences recherchées se trouvent dans le dernier tiers de la chanson où leur fréquence augmente rapidement.

L'axe vertical nous apporte les données sur la fréquence ; au total, le texte cible contient cinq occurrences du mot « kámo », le texte source comprend neuf occurrences de ses équivalents français.

Un clic sur le bouton *concordance des formes sélectionnées* dans l'onglet ***dic*** nous permet de visualiser la concordance de mots sélectionnés, soit sous forme du contexte-partie, soit sous forme de la tri-concordance. On peut apercevoir les différences sur les figures ci-dessous. Tandis que le premier type d'affichage visualise toujours

une partie de la phase contenant un mot sélectionné, le deuxième nous montre la même partie divisée en trois colonnes ce qui augmente la clarté de la visualisation.

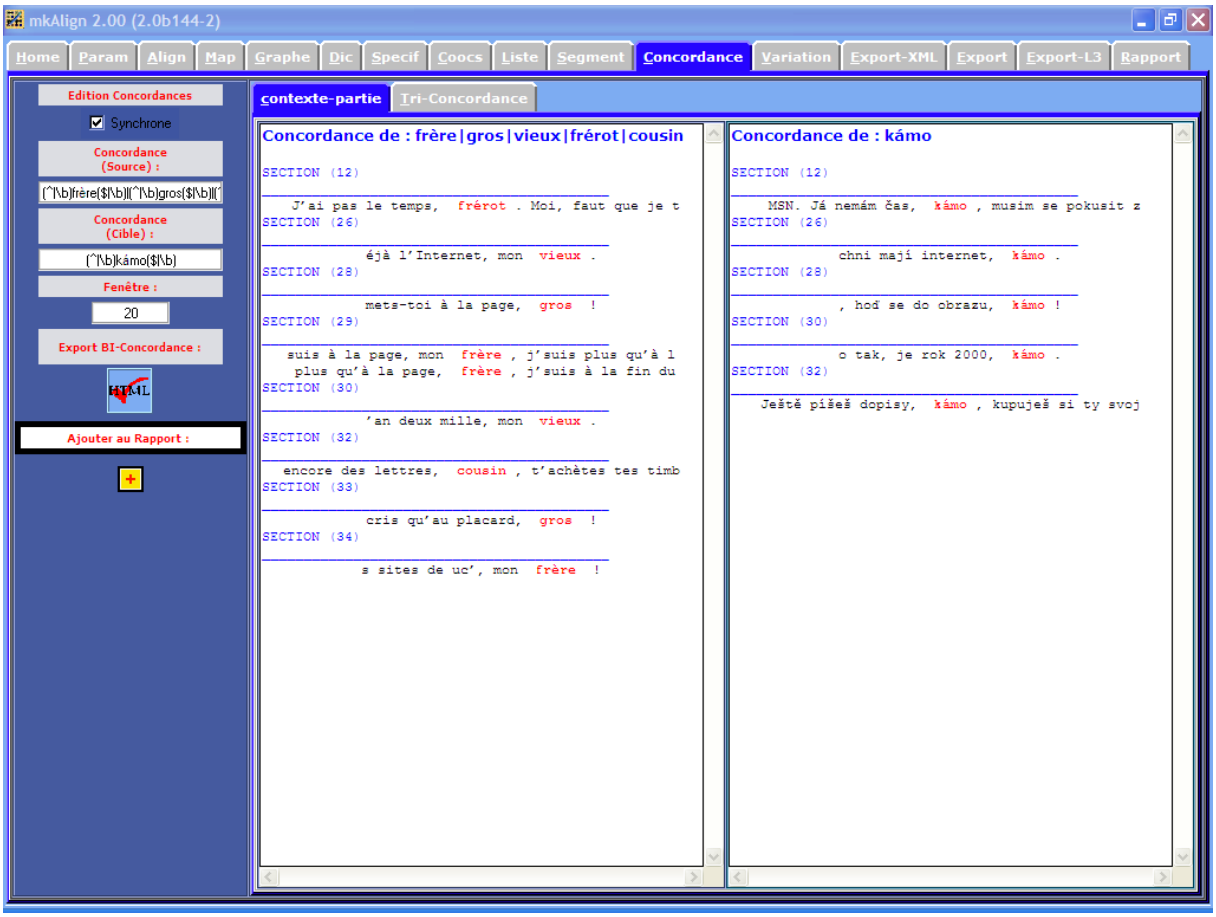


Figure n°6 : Concordance des formes sélectionnées (contexte-partie)

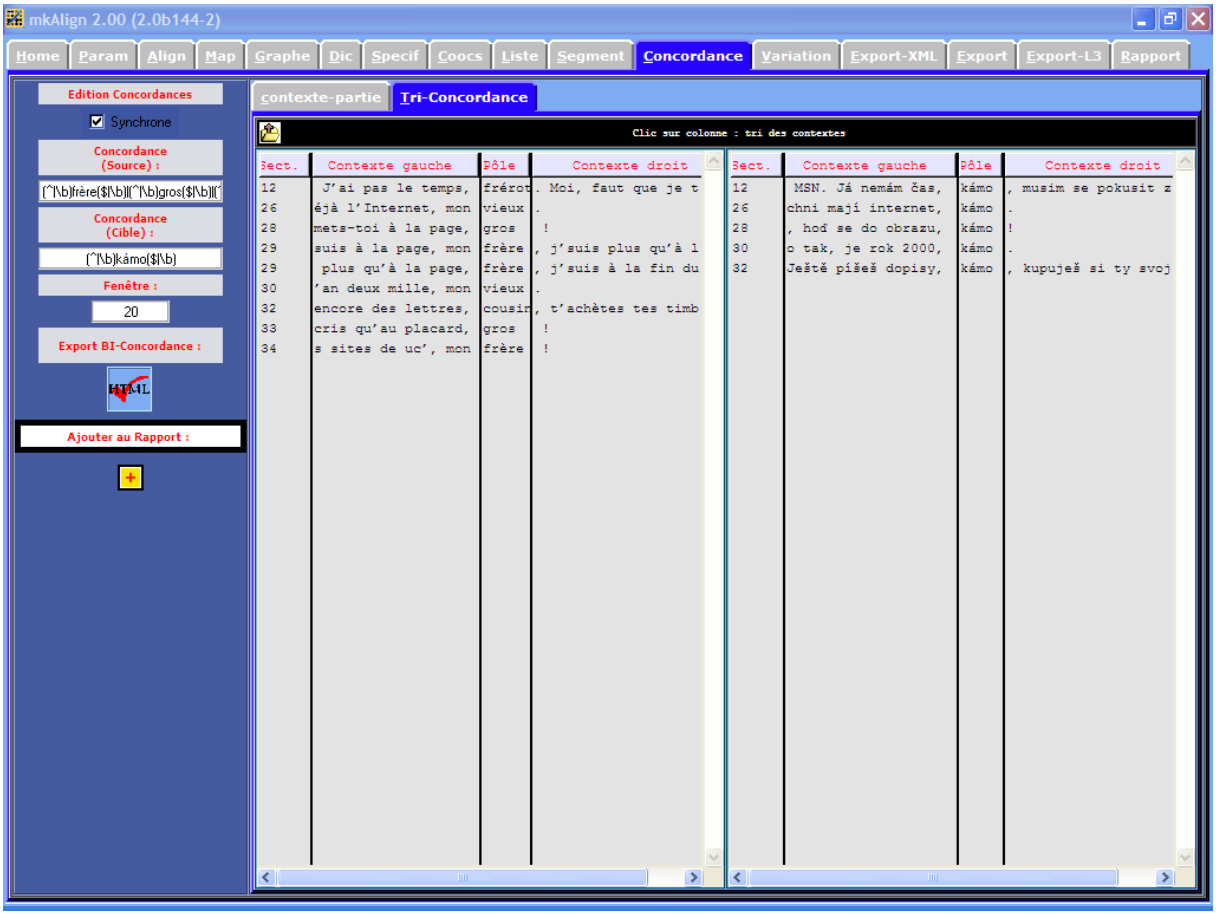


Figure n°7 : Concordance des formes sélectionnées (Tri-Concordance)

Les fonctions du *MkAlign* décrites ci-dessus ont été conçues plus particulièrement pour qu'on puisse comparer un texte original avec sa traduction. En outre, le programme permet de comparer deux versions d'un même texte pour découvrir ainsi les différences. Ce processus est réalisable par l'intermédiaire de l'onglet ***variation*** représentant le sujet d'un chapitre suivant.

3 PARTIE PRATIQUE : VARIATION

L'onglet ***variation*** offre à l'utilisateur la possibilité de comparer automatiquement deux versions d'un texte et de découvrir ainsi la mise à jour de la variation entre ces deux textes. Le processus de repérage de la variation se déroule en deux étapes. La première consiste en l'alignement des deux versions d'une manière déjà mentionnée dans le sous-chapitre 2.1.1. Nous enregistrerons donc les textes d'après cette méthode et puis, nous ouvrirons l'onglet ***variation*** en accédant à la deuxième étape. La figure n°8 représente la barre d'outils disponibles dans ce volet.

Figure n°8 : Variation – la barre d'outils (Source : FLEURY, Serge. *Op. cit.*, p. 63)

D'abord, nous choisirons le paramétrage d'un calcul en cochant sur l'un de trois carrés : mot, ligne ou caractère, puis nous lancerons l'opération par un clic gauche sur le bouton  . L'objectif de ce processus est de repérer « la variation par coloration des ajouts, suppressions, modifications »[69]. Grâce à cette fonction, on est capable de comparer une version « pochette » avec une version « son » d'une même chanson rap conformément à l'objectif de ce mémoire.

À titre d'exemple, nous utilisons les textes de la chanson « Salope.com », composée et interprétée par le rappeur Kool Shen sur son album Crise de conscience, pour montrer le processus dans son entier. Premièrement, nous ajusterons les deux versions comme décrit dans le sous-chapitre 0., mais nous garderons les parties qualifiées comme excessives (« Intro », « Couplet 1 », etc., dans notre cas) pour pouvoir analyser les résultats d'une manière plus effective ensuite. Successivement, nous contrôlerons la ponctuation et les caractères cachés. Pour faciliter le traitement des textes, nous choisirons le mode général dans le volet ***align*** du programme *MkAlign* et un prétraitement opérant « comme un segmenteur ». Nous éviterons ainsi la nécessité de compléter le caractère # dans les textes. C'est la ponctuation qui y servira de segmenteur. Deuxièmement, nous enregistrerons les textes dans le programme et nous modifierons ce « bi-texte » créé afin les cellules alignées se correspondent. Puis,

[69] S. FLEURY, *MkAlign (version 2.0) : Manuel...*, *op. cit.*, p. 64.

nous ouvrirons le volet ***variation*** où nous lancerons le processus de comparaison. On peut apercevoir le résultat de cette action sur la figure n°9[70].

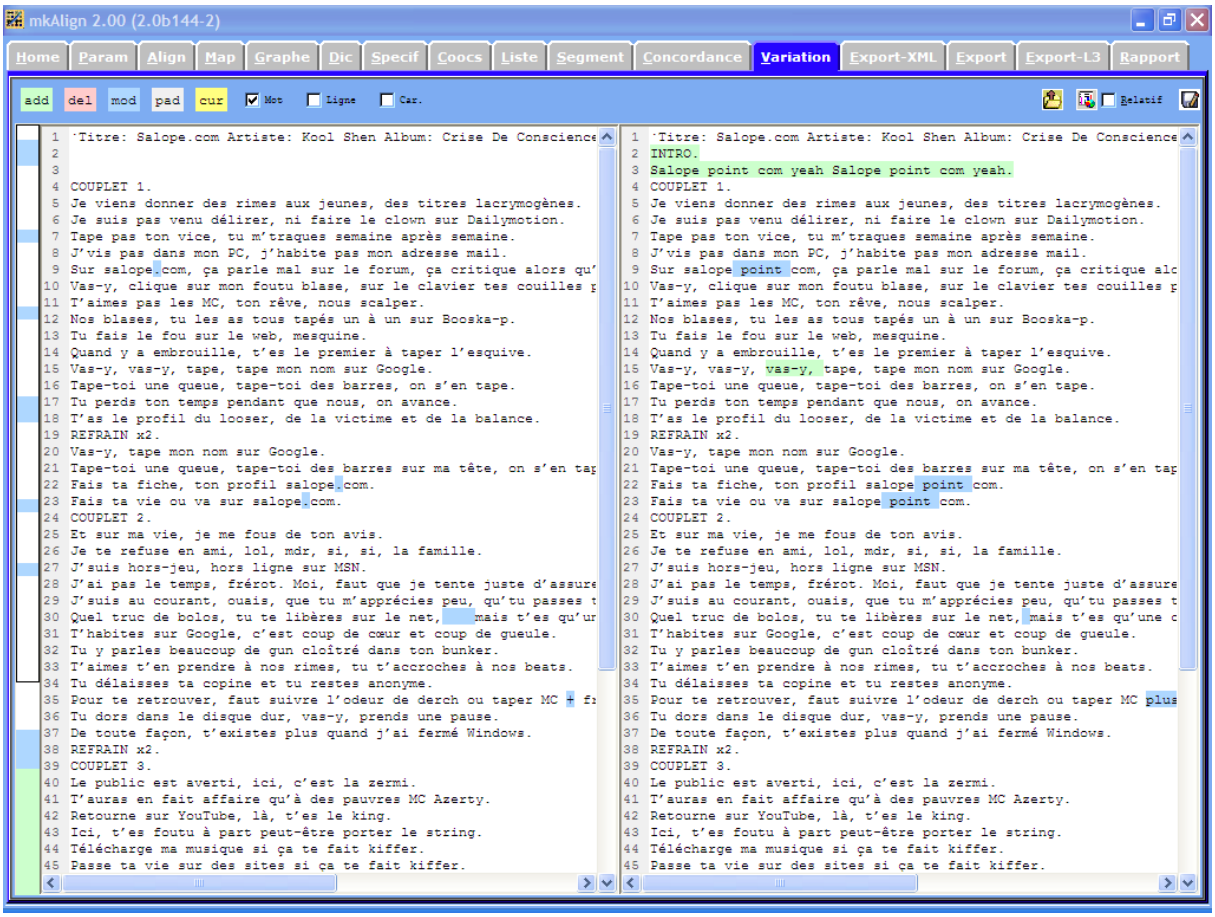


Figure n°9 : Variation – Salope.com (Kool Shen)

De cette façon, nous analyserons l'échantillon de cinquante chansons rap, la chanson « Salope.com » compris.

3.1 Aspects méthodologiques

Notre échantillon est composé de chansons trouvées sur le site du RapCor[71]. Le choix a été réduit par la disponibilité des chansons déjà prêtes à être traitées par logiciel TXM. Actuellement, on y trouve sept cents neuf chansons répondant à cette condition. Dans notre échantillon, nous avons incorporé les chansons dont les interprètes ont au moins cinq chansons dans ce sous-corpus. Il s'ensuit que chaque cinquième chanson

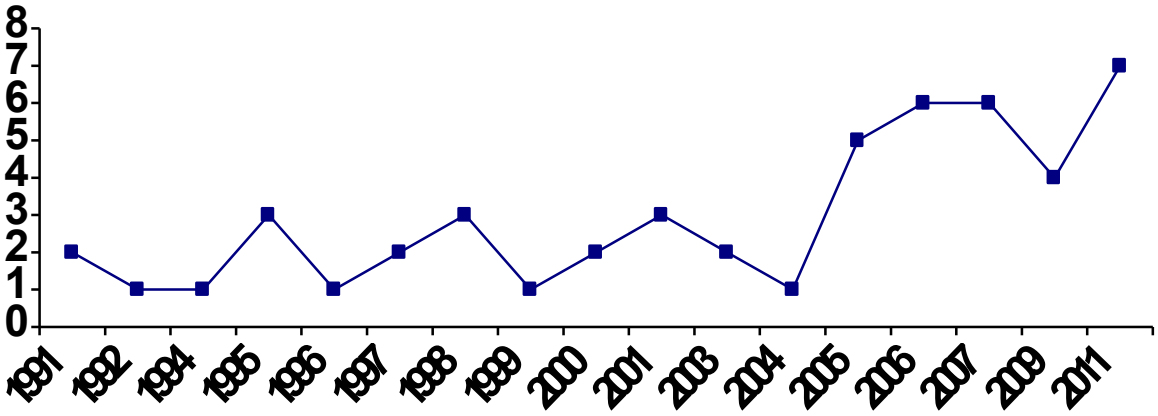
[70] La visualisation de toute la comparaison des deux versions (version « pochette », version « son ») est disponible sur le CD joint (sous nom « Export variation – Kool Shen (Salope.com).
[71] Le corpus se trouve sur ce lien : <https://is.muni.cz/auth/do/phil/Pracoviste/URJL/rapcor/data_zaloha/>.

de cet interprète, éventuellement chaque dixième chanson[72], est incluse dans notre échantillon. Néanmoins cette sélection n'est pas aussi stricte. De temps en temps, la cinquième chanson d'un certain interprète est manquante. C'est pour cette raison que nous l'avons remplacée par la chanson qui était la plus proche de cette « cinquième » chanson. De cette raison, on peut dire que la sélection des chansons a été aléatoire. Le corpus est composé de deux versions d'une chanson, de sa version « pochette » et de sa version « son ». La liste des chansons analysées se trouve en annexe.

Cherchant à trouver une ou plusieurs règles dans les différences des deux versions texte et son, nous traiterons ce sujet à partir de deux points de vue. D'abord, nous nous orienterons vers la structure de ces deux versions, c'est-à-dire vers le côté formel des chansons, puis nous examinerons cette problématique du côté textuel.

3.2 Chansons analysées

Pour que nous puissions passer à la comparaison des versions texte et son des chansons choisies, nous nous occuperons d'abord de l'analyse générale de ces chansons de rap. Nous présenterons des données statistiques, notamment la représentation des différentes dates de sortie des diverses chansons comprises dans notre échantillon. Puis, nous nous orienterons vers une analyse plus détaillée et vers la détection des différences entre les versions texte et son.



Graphique n°1 : Échantillon des chansons rap – représentation des dates de sortie des chansons

[72] Si ce corpus contient plus de quinze chansons d'un interprète, nous avons inclus aussi chaque dixième chanson de cet interprète, cela veut dire la chanson n°5, puis n°15, n°25 etc.

La figure n°10 nous montre la représentation des différentes années de sortie des chansons rap analysées. Les chansons qui sont sorties avant l'année 2000 forment 28 % de l'échantillon. La plupart des chansons analysées sont sorties après le tournant du millénaire. Comme nous l'avons déjà mentionné dans l'introduction, le projet est lié au cours Langage des jeunes ce qui est aussi la raison pour laquelle des chansons récentes prédominent. Nous puisons des renseignements sur le langage, entre autres, des chansons rap et étant donné que la recherche est orientée vers le langage actuel, nous nous efforçons d'utiliser principalement les chansons de la date de création la plus récente.

> Le rap français est la version française du rap américain qui a vu le jour avec l'arrivée de groupes comme IAM, Suprême NTM, Assassin ou MC Solaar. Tout en restant continuellement inspiré par les rappeurs d'outre-Atlantique, le rap français élabore progressivement sa propre personnalité, oscillant entre revendications sociopolitiques, messages positifs ou festifs et tentation commerciale.[73]

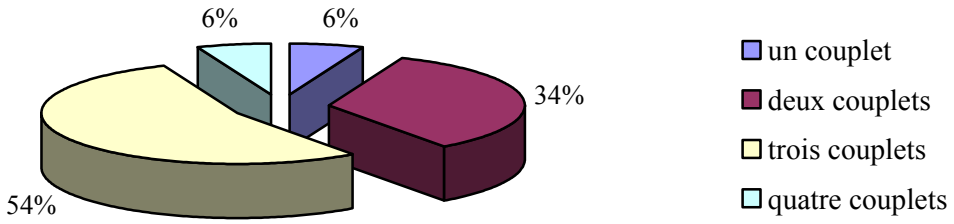
En ce qui concerne les interprètes, la plupart d'entre eux sont représentés par une ou deux chansons. Nous nous sommes fixée comme contrainte de choisir des chansons de différents albums, les dates de sortie ne devraient pas être influencées par ce fait. Deux interprètes, IAM et MC Solaar, sont l'exception. Quatre chansons de ces deux interprètes figurent sur notre liste de chansons choisies. La librairie des chansons rap contient beaucoup plus de chansons de ces deux rappeurs en comparaison avec les autres interprètes de notre échantillon. Il faut mentionner que le choix des interprètes n'a pas été si strict qu'il peut le sembler. Même si nous nous sommes forcée d'observer la règle décrite plus haut, nous n'avons pas réussi à éviter les exceptions. À cause d'une limitation numérique, nous étions obligée d'omettre par exemple plusieurs d'interprètes ; c'est pourquoi notre liste ne comporte pas leur chanson. Le nombre des chansons comprises est limité à cinquante sur un nombre total de 709 chansons, traitées à l'aide du logiciel TXM, se trouvant dans la librairie. Cet échantillon ne représente que 7 % du nombre total, il ne peut donc pas être considéré d'être doté d'une valeur de référence. Néanmoins, on peut supposer que les résultats constatés ci-dessous peuvent être applicables pour un échantillon plus étendu que le nôtre, parce que les chansons rap, comme tous les genres spécifiques, partagent certains éléments. Entre

[73] *Le Rap français (Histoire et définition)* [en ligne]. Page consultée le 2013-10-15. Disponible à l'adresse : <http://www.rap2banlieue.com/rap-francais/>.

outres, il faut souligner ceux qui contribuent au dévéloppement d'une musicalité de ce type de chansons.

> [...] Toutes ces transformations phonétique, ruptures soudaines, syncopes, élision systématique de la consonne finale du gérodif, synocope, élission systématique de la consonne finale du gérondif, rendent plus fluide le récit du rappeur, permettent une plus grande densité de pieds dans le même vers et, au final, font que le rap se récite en moyenn quatre fois plus rapidement que n'importe quel autre style musical.[74]

Notre échantillon contient les chansons rap d'une structure différente. Le graphique suivant montre un pourcentage des diverses représentations structurelles, des chansons à un couplet aux chansons à quatre couplets :



Graphique n°2 : Échantillon des chansons rap – représentation structurelle

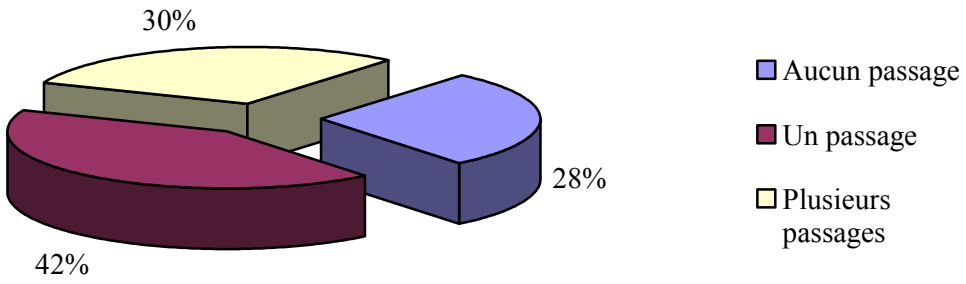
Comme on peut le voir sur la figure n°11, les chansons à trois couplets prédominent. Il s'agit de la structure typique pour ce type de chansons. Néanmoins, les chansons ne sont pas seulement divisées en couplets séparés par des refrains, mais quelques-unes sont munies aussi d'intro, d'interlude ou d'outro. Ces parties sont ajoutées par les rappeurs le plus souvent pendant l'interprétation de la chanson et elles ne figurent que dans les versions son.

3.2.1 Côté formel des chansons analysées

Dans ce mémoire mineur, le côté formel est évoqué d'un point de vue structurel. Nous analysons une structure des différentes chansons et nous comparons leur version texte à la version son. Généralement, la chanson peut être composée de plusieurs parties, soit une intro, des couplets, un interlude, un refrain, une outro. L'intro ou l'introduction est

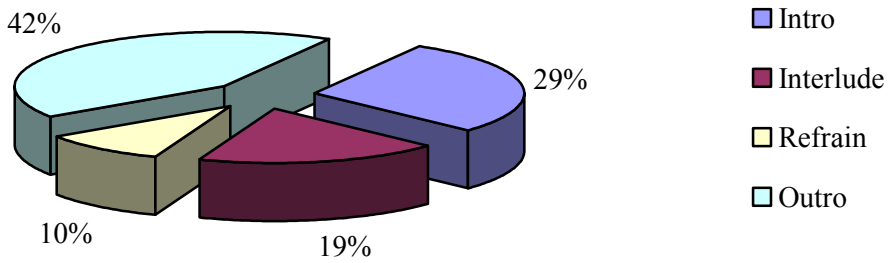
[74] Janette BECKMAN et B. SANDLER, *RAP! : Portraits and lyrics of a generation of Black Rockers*, New York, St Martin's Press, 1991, cité par David DIALLO, « La musique rap comme forme de résistane ? », *Revue de recherche en civilisation américaine*, 2009, 1, p. 8.

le passage introduisant la chanson. Il s'agit le plus souvent d'un ou plusieurs mots sous forme d'une exclamation exprimant des émotions. L'outro représente l'opposé de l'intro, c'est-à-dire la conclusion d'une chanson. Ces deux passages sont omis le plus souvent dans une version « pochette » et ils sont ajoutés à la version « son » additionnellement, lors de l'interprétation. Les graphiques ci-dessous indiquent les différences entre les versions « pochette » et les versions « son » de notre échantillon.



Graphique n°3 : Échantillon des chansons rap – passages ajoutés

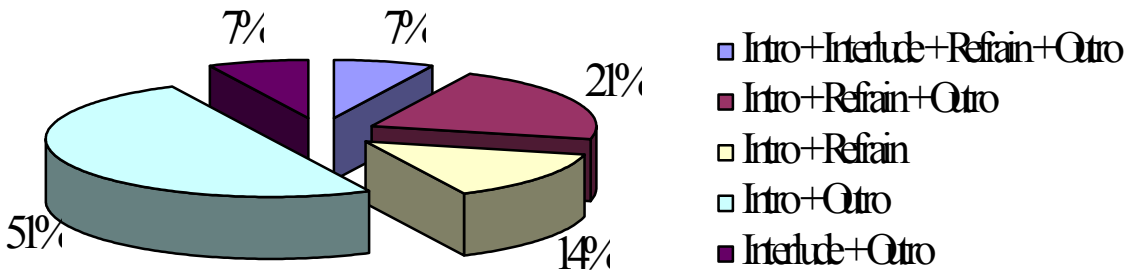
La figure n°12 nous montre une représentation des diverses modifications structurelles en indiquant le nombre de passages ajoutés. Ce ne sont que 28 % des chansons examinées qui sont restées les mêmes du point de vue structurel au cours de leur interprétation. Les autres ont été élargies. Les graphiques suivants nous en donnent une description plus détaillée. La première analyse les chansons comprenant un passage ajouté, soit l'onglet de 42 % du graphique précédent.



Graphique n°4 : Analyse des chansons à un passage ajouté

Comme on peut le voir sur la figure n°13, l'outro prédomine sur tous les autres passages ajoutés aux versions son. En chiffres réels, il s'agit de neuf chansons sur vingt et une chansons modifiées de cette manière-ci. L'outro représente aussi une partie importante dans les chansons à plusieurs passages ajoutés. Le graphique suivant

représente une preuve de cette affirmation. Le graphique n°14 analyse l'onglet de 30 % sur la figure n°12.



Graphique n°5 : Analyse des chansons à plusieurs passages ajoutés

En analysant le graphique n°13 et le graphique n°14, on peut constater que les passages ajoutés le plus souvent sont l'intro et l'outro, soit le commencement et la fin de la chanson. On peut en déduire que la plupart des interprètes préfèrent ne pas modifier radicalement la structure de leurs chansons. Néanmoins, il n'y a qu'un faible nombre de chansons qui restent les mêmes, sans la moindre modification. Nous traiterons les différents passages ajoutés dans le paragraphe suivant.

3.2.1.1 Traits caractéristiques pour les intros examinés

Après l'analyse de cinquante chansons, nous pouvons constater que 36 % d'entre elles sont munies d'une intro. En comparaison avec les autres passages ajoutés, l'introduction représente généralement le passage le plus court entre les passages qui n'apparaissent pas dans les versions « pochette ». Son contenu peut être varié. Les exemples de son utilisation sont cités ci-dessous, la liste exhaustive se trouve en annexe.

a) l'intro servant souvent à introduire la chanson et à faire allusion à son contenu.

ARS05 – La vie, ça tient à rien. Il suffit d'une balle, il suffit d'une balle.
KEN25 – Petit frère a déserté le terrain de jeu...
KER05 – Si c'était à refaire, Kery James, deux mille un, il y a pas de couleur pour pleurer, je dis, il y a pas d'amour pour aimer.

b) l'intro dédiant la chanson à une personne ou à un groupe spécifique

DIA15 – C'est pour toutes mes sœurs de France ou d'ailleurs, c'est pour vous, ça !
FLY05 – Eh yo, c'est Flynt. Dédicace aux rageurs avec un majeur à chaque doigt (x2)

c) l'intro ne contenant que le titre de la chanson ou le nom de l'interprète

FAY05 – Vraies liaisons et lesions.

KSH03 – Salope point com, yeah. Salope point com, yeah.

MIC01 – Colonel Reyel (x2). Mister You. Mets-toi à l'aise. Mister You. Toi, toi, toi, mets-toi à l'aise.

d) l'intro se présentant sous forme d'une exclamation, d'un encouragement

DON05 – Ho ! Ho ! Enculé ! Docteur ! Docteur ! Laisse-moi sortir ! Hoo ! Hoooooo !

MCS45 – Allez, allez !

3.2.1.2 Traits caractéristiques pour les interludes examinés

En comparaison avec les autres passages ajoutés, l'interlude n'apparaît que dans sept chansons, soit 14 % de tout l'échantillon examiné. Le dictionnaire Linternaute.com définit l'interlude comme un « petit divertissement entre deux émissions ou deux parties d'un spectacle »[75]. Cette définition, un peu modifiée, peut être appliquée aussi aux chansons de rap où l'interlude représente un divertissement entre deux couplets. Sa forme est variée. Par exemple, l'interlude de la chanson nommée « Le poison rouge » (PLD02) contient l'explication de ce terme sous forme d'un dialogue où Disiz La Peste, l'interprète, appelle Treize à raconter l'histoire lié au poison rouge. À travers la chanson « On est encore là », le groupe NTM lutte contre la censure en disant que ses membres sont toujours prêts à dire la vérité, même s'ils sont obligés de faire face aux condamnations et peines possibles. Dans l'interlude, l'interprète mentionne le procès mené auprès de la Cour d'appel d'Aix-en-Provence contre ce groupe et il précise ainsi la situation, même si jusqu'à ce moment, il ne parle que généralement. On peut constater que le rappeur rapproche ainsi l'audience de la réalité du groupe et il précise le sujet de la chanson. Par contre, l'interlude de la chanson « Plus rien ne m'étonne » de l'interprète Orelsan ne contient qu'une modification du refrain. Il s'ensuit que nous ne sommes pas capables de formuler une définition de l'interlude sur la base de l'échantillon de sept chansons.

[75] Linternaute.com. *Interlude : définition et synonymes du mot interlude dans le dictionnaire* [en ligne]. Page consultée le 2013-03-02. Disponible à l'adresse : <http://www.linternaute.com/dictionnaire/fr/definition/interlude/>.

3.1.2.3 Traits caractéristiques pour les refrains examinés

Le refrain a été ajouté à dix chansons, soit à 20 % de tout l'échantillon. Nous n'avons pas pris en compte les refrains répétés par exemple deux fois à la suite dans la version « son », même si le refrain n'a été mentionné qu'une fois dans la version « pochette ».

En général, ces dix chansons peuvent être divisées en deux groupes. Le premier, composé de trois chansons, contient en version « son » le refrain dont le texte ne figure absolument pas dans la version « pochette ». Au contraire, le refrain des chansons se trouvant dans le deuxième groupe est répété plusieurs fois contrairement à la version « pochette », il se trouve à plusieurs places dans la chanson. En ce cas, le refrain additionnel a été ajouté le plus souvent après le dernier couplet. Dans quatre cas sur dix, il n'y a qu'un refrain additionnel en version son. Les trois autres chansons[76] sont munies de plusieurs occurrences de cette partie.

Néanmoins, l'écriture du refrain n'est pas si nette. Sur l'exemple de la chanson « Les filles sont belles », nous pouvons constater qu'il y a au moins deux manières de mettre en page le texte de la chanson. La première manière suit une succession des différentes parties de la chanson en les nommant les unes après les autres. L'autre manière ne mentionne que la première occurrence du refrain en supposant comme évident que les autres suivront entre les couplets. Cela peut être la raison pour laquelle la chanson LSP06, KDD05 ou KEN25 ne contient qu'une mention du refrain en version « pochette ». Néanmoins, ce n'est pas le cas de la chanson BOO05, CAN05[77], MAF05[78], SEP05. Dans leurs versions « pochette », le refrain est mentionné plusieurs fois (le refrain additionnel est ajouté le plus souvent après le dernier couplet). En ce qui concerne les chansons ALI05, ASS05 et MCS45, leurs refrains ne figurent pas du tout sur la version « pochette ».

3.1.2.4 Traits caractéristiques pour les outros examinés

En totalité, vingt et une chansons sont munies de l'outro, soit 42 % de l'échantillon examiné, ce qui fait de l'outro la partie la plus souvent ajoutée. Il est possible de diviser l'outro en plusieurs types. Dans notre cas, nous pouvons constater qu'il y a deux

[76] Il s'agit de KDD05, KEN25 et LSP06.
[77] C'est un cas exceptionnel, parce qu'il s'agit de la chanson à deux couplets.
[78] Le même cas comme CAN05.

groupes principaux d'outros. Le premier groupe contient les outros de quatre chansons, soit ARS05, CAN05, DIA07 et SNS01 qui ne sont représentées que par des exclamations comme « négro, négro » ou « Casanière ». Ces exclamations peuvent être répétées plusieurs fois et sous formes variées. C'est le cas par exemple de DIA07. L'interprète scande différentes variantes de son nom et à la fin, elle ajoute « Ouais, grosse ! ».

Les textes des outros du deuxième groupe sont plus développés et en général, ils délivrent un message ou une dédicace. Par exemple, la chanson de l'interprète Kery James, KER15, contient une dédicace à plusieurs personnes.[79] Les chansons comme IAM05, KDD05 ou ORE07 sont munies d'outros répondant à leur interlude ou à leur refrain modifié. Le message et en même temps la conclusion est représenté par l'outro de la chanson ASS15 dont le texte est le suivant : « Le business illégal est légal quand il sert le patrimoine national ». L'outro de la chanson KER05 lutte contre le racisme comme toute la chanson et il est composé de diverses parties de la chanson. L'outro de la chanson JAD05 relie l'interlude et le nom de l'interprète. Pour cette raison, il est difficile de diviser l'outro en plusieurs types en les décomposant d'une manière stricte. Il s'agit souvent d'une combinaison de plusieurs éléments ; l'outro peut contenir le nom de l'interprète scandé, soit l'outro défini dans ce mémoire mineur correspondant au premier type, et en même temps une modification de l'interlude, donc appartenant au deuxième type.

L'outro de quelques chansons – par ex. KSH03 – est beaucoup plus long en comparaison avec les outros des autres chansons dans l'échantillon examiné. Généralement, l'interprète développe le sujet de la chanson. Dans le cas de la chanson KSH03, l'interprète mène un dialogue au sujet de l'Internet, le thème de la chanson, avec son ami. Ce qui est difficile dans ces outros développés est de transcrire la version son sous forme écrite. Habituellement, l'interprète utilise beaucoup de mots familiaux et il ne prend pas soin de la prononciation soutenue. De cette raison, la transcription n'est pas souvent exacte.

Les parties omises – peu importe s'il s'agit de l'intro, de l'interlude ou de l'outro – sont de caractères différents, mais leurs contenus ne sont pas si violents ou impertinents pour

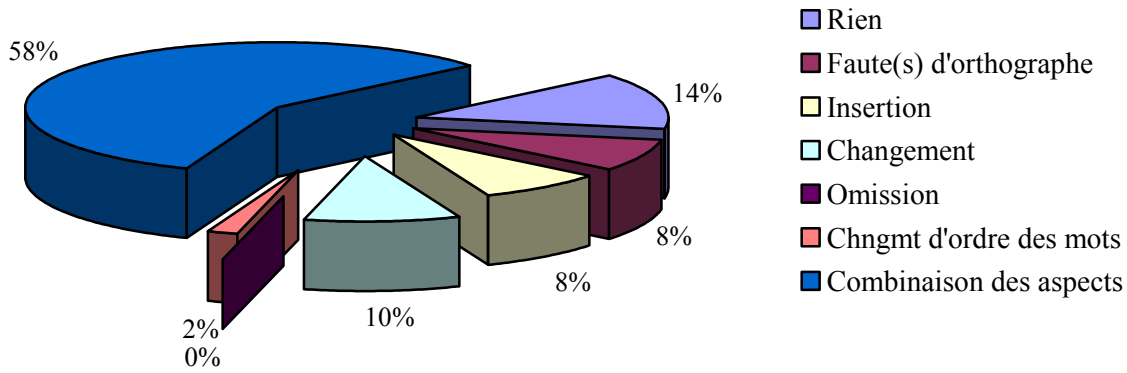
[79] « Je dédie ce morceau à tous ceux qui peu à peu perdent confiance en la confiance... Je dédie ce morceau à ma ptite sœur, Casandra, à mot ptit frère, Kévin, et à ma mère... »

qu'ils puissent causer l'interdiction de la publication de la chanson. En général, on peut constater que ces parties servent à attirer l'attention du public ou à accentuer le thème de la chanson.

3.2.2 Côté textuel des chansons analysées

En analysant le côté textuel, nous avons pris en considération les aspects suivants : les fautes d'orthographe, l'insertion d'un mot ou d'une phrase, le changement d'un mot, l'omission d'un mot ou d'une phrase et le changement d'ordre des mots dans une phrase. De plus, nous mentionnerons la répétition additionnelle de mots ou de phrases dans les versions « pochette » ou « son ». Nous traiterons d'abord ces aspects d'une manière générale et puis, nous les décrirons de manière plus détaillée dans les différents sous-chapitres.

Le graphique n°15 visualise la représentation de ces aspects dans l'échantillon de cinquante chansons.



Graphique n°6 : Analyse du côté textuel

Il n'y a que sept chansons, soit 14 % du total, dont la version « pochette » et la version « son » sont les mêmes au niveau textuel. L'échantillon de quatorze chansons, soit 28 % d'un total, ne contient qu'un des aspects nommés ci-dessus. Le graphique n°15 spécifie la répartition de ces aspects. Il s'agit des catégories suivantes : les fautes d'orthographe, l'insertion d'un mot ou d'une phrase, le changement ou la modification d'un mot, l'omission d'un mot ou d'une phrase et le changement d'ordre des mots. Le reste des cinquante chansons examinées, soit vingt neuf chansons ou 58 %, combine plusieurs aspects en même temps. Après cette examination, nous pouvons constater que

le minimum des chansons (14 %) reste la même dans les deux versions, tandis que la plupart (58 %) des versions « pochette » et « son » diffèrent en plusieurs aspects.

3.2.2.1 Typologie des fautes d'orthographe

Le Petit Robert[80] définit le mot « orthographe » comme une « manière d'écrire un mot qui est considérée comme la seule correcte ». Par contre, le mot « faute » signifie un « manquement à une règle, à un principe »[81]. Au total, vingt-trois chansons, soit presque la moitié de l'échantillon, contiennent au moins une faute d'orthographe, mais seulement certaines de ces fautes ont une influence sur la prononciation.

3.2.2.1.2 Fautes d'orthographe sans influence sur la prononciation

Les fautes sans influence sur la prononciation peuvent être divisées en deux groupes. Les fautes dans les deux groupes sont généralement causées par l'ignorance d'une forme correcte. Le premier groupe (A) ne contient que les fautes d'orthographe ne provoquant pas de changement de sens du mot touché. Les fautes dans le deuxième groupe (B) changent le sens du mot.

A) Les fautes d'orthographe sans influence sur la prononciation et sans changement de sens sont de plusieurs types. Le plus souvent, il s'agit de faute dans les mots contenant une lettre redoublée, de faute de ponctuation, de conjugaison, de fautes liées à l'omission ou à l'addition d'une lettre. Vous trouverez les exemples ci-dessous.

- **Faute liée à la lettre redoublée**

CGT[82]	**ST**[83]	**P**[84]	**S**[85]	**CC**[86]
le nom	1	Samara	Samarra	LAK01
l'adjectif qualificatif	1	internationnaux	internationaux	ASS15

[80] Le Nouveau Petit Robert de la langue française 2008, p. 1763.
[81] *Ibid.*, p. 1018.
[82] Le sigle « CGT » signifie la « catégorie grammaticale touchée ».
[83] Le sigle « ST » signifie la « somme totale » des exemples provenant de la même catégorie.
[84] Le sigle « P » signifie la version « pochette » de la chanson.
[85] Le sigle « S » signifie la version « son » de la chanson.
[86] Le sigle « CC » signifie le « code de la chanson ».

le verbe	1	ennivrée	enivrée	KEN03

Tableau n°1 : Faute liée à la lettre redoublée (fautes d'orthographe sans influence sur la prononciation)

- **Faute de ponctuation**

CGT	ST	P	S	CC
le nom	6	coté	côté	ARS05
		cocaîne	cocaïne	ASS15
		problémes	problèmes	
		heroîne	héroïne	
		envies de peze	envies de pèze	DIA15
		l'hopital	l'hôpital	DLP02
le verbe	4	engrèner	engrener	ARS05
		controlés	contrôlés	ASS15
		brulé	brûlé	DLP02
		réapparaîs	réapparais	LSP06
l'adjectif qualificafif	2	fièr	fier	ARS05
		moyen-ageux	moyenâgeux	BOO05

Tableau n°2 : Faute de ponctuation (fautes d'orthographe sans influence sur la prononciation)

- **Faute de conjugaison**

MVT[87]	ST	P	S	CC
l'impératif	7	trouves moi	trouve-moi	CAN05
		rapportes moi	rapporte-moi	
		Joues pas [...]	Joue pas [...]	MAF05
		Avoues [...]	Avoue [...]	
		Soit pas bloqué [...]	Sois pas bloqué [...]	
		Aller, viens, on y go [...]	Allez, viens, on y go [...]	MIC01
		tu veux parler, vas-y, parles bien sinon [...]	tu veux parler, vas-y, parle bien sinon [...]	SNS01
l'indicatif	2	si tu coopère [...]	si tu coopères [...]	MAF05
		j't'en parles [...]	j't'en parle [...]	OXM15
l'infinitif	2	petit soldat aimerait posé l'épée	petit soldat aimerait poser l'épée	KEN25

[87] Le sigle « MVT » signifie le « mode de verbe touché ».

		[...] laisse les foncedé [...]	[...] laisse les fonceder [...]	LAK01
le futur proche	1	on vas te tuer	on va te tuer	OXM15
le subjonctif	1	pour que t'évite	pour que t'évites	DON05

Tableau n°3 : Faute de conjugaison (fautes d'orthographe sans influence sur la prononciation)

- **Faute liée à l'addition ou à l'omission d'une lettre**

TF[88]	ST	P	S	CC
le pluriel	5	un paquet de cigarette	un paquet de cigarettes[89]	ASS15
		le coup malgrès leurs...	le coup malgré leurs...	CAN05
		autant d'auréole	autant d'auréoles	LAK01
		pleins, pleins d'mecs	plein, plein d'mecs	MIC01
		Sarah et sa copine me donnent leur téléphone [...]	Sarah et sa copine me donnent leurs téléphones[...]	
la terminaison	2	c'pays d'clébars	c'pays d'clébards	BOO05
		jeune beurre !	jeune beur !	LAL07

Tableau n°4 : Faute liée à l'addition ou à l'omission d'une lettre (fautes sans influence sur la prononciation)

B) Le deuxième groupe rapporte les fautes changeant le sens du mot, même si cette faute n'est pas perceptible dans la version sonore. Comme exemple, nous pouvons citer la chanson ALI05 où la locution prépositionnelle « quant à » est remplacée incorrectement par la conjonction « quand » dans la version « pochette ».[90] Cette faute se trouve aussi dans la version « pochette » de la chanson ASS15. Par contre, le rappeur y utilise « quant » au lieu de « quand ».[91] Dans quelques cas, il est discutable si la faute a été commise sur la version « pochette ». Par exemple, dans la version « pochette » de la chanson ASS15, nous trouvons la préposition « à » qui est remplacée par le verbe « avoir » à la troisième personne du singulier dans

[88] Le sigle « TF » signifie le « type de la faute ».
[89] Cela peut être considérée comme une faute commise dans une version son.
[90] ALI05. Version « pochette » : « J'ai ouvert les yeux quand à mon sort... » X version « son » : « J'ai ouvert les yeux quant à mon sort.... ».
[91] ASS15. Version « pochette » : « Tous les business sont légaux quant ils sont controlés... » X version « son » : « Tous les business sont légaux quand ils sont contrôlés... ».

la version son.[92] Cependant les fautes de ce type n'influencent pas la prononciation et il n'est pas possible de les noter en version son. Ce changement de sens n'est pas donc perceptible lors de la performance de la chanson.

3.2.2.1.2 Fautes d'orthographe avec l'influence sur la prononciation

Les fautes d'orthographe ayant une influence sur la prononciation sont également de plusieurs types, même si leur liste n'est pas aussi exhaustive que dans le cas précédent. Le premier groupe contient les fautes liées à la lettre redoublée, le deuxième comprend les fautes liées à l'addition ou l'omission d'une lettre ou d'une syllabe. De plus, il y a aussi deux groupes qui ne se trouvent pas dans la catégorie I, les fautes liées à la modification d'une lettre dans un mot et les fautes liées à l'élision. Des exemples sont mentionnés ci-dessous.

- **Faute liée à la lettre redoublée**

CGT	ST	P	S	CC
le nom	3	des tranquilisants	des tranquillisants	ASS15
		les malettes	les mallettes	FFA05
		du gallon	du galon	LAL07

Tableau n°5 : Faute liée à la lettre redoublée (fautes d'orthographe avec l'influence sur la prononciation)

Ces fautes n'influencent la prononciation que théoriquement. Les interprètes sont informés sur la prononciation correcte des mots nommés, même si leur forme écrite ne répond pas à l'orthographe approuvée et selon ses règles, la prononciation devrait être différente.

- **Faute liée à l'addition ou à l'omission d'une lettre ou d'une syllabe**

TF	ST	P	S	CC
la terminaison	3	la misécorde	la miséricorde	AKH06
		mama	maman	LAL07
		une vieille	une vielle[93]	ORE07

[92] ASS15. Version « pochette » : « Mais attention, chacun à sa place, sniffe de la coke dans le ghetto, vends du teuchi par kilo... » X version « son » : « Mais attention, chacun a sa place, sniffe de la coke dans le ghetto... ».

[93] Dans ce cas, la faute est plutôt en version son. La phrase entière est la suivant : J'viens d'voir une vieille (en P)/vielle (en S) faire une crise cardiaque, premier réflexe, j'ai twitté.

le pluriel	2	des p'tits larcins	de p'tits larcins	LAL07
		j'te parle de faits divers [...]	j'te parle des faits divers [...]	SNS01
le radical	1	Rockfeller	Rockefeller	KOM05

Tableau n°6 : Faute liée à l'addition ou à l'omission d'une lettre ou d'une syllabe (fautes d'orthographe avec l'influence sur la prononciation)

- **Faute liée à la modification d'une lettre**

TF	**ST**	**P**	**S**	**CC**
la terminaison	1	mon côté tacitume	mon côté taciturne	FAY05
le radical + la terminaison	1	falhen	farhan	LAK01

Tableau n°7 : Faute liée à l'addition ou à la modification d'une lettre (fautes d'orthographe avec l'influence sur la prononciation)

- **Faute liée à l'élision**

CGT	**ST**	**P**	**S**	**CC**
la conjonction de subordination	1	si il ne faudrait pas [...]	s'il ne faudrait pas [...]	ASS15
le pronom relatif	1	[...] mecs qu'ont la dalle !	[...] mecs qui ont la dalle !	MIC01

Tableau n°8 : Faute liée à l'élision (fautes d'orthographe avec l'influence sur la prononciation)

Même si toutes les fautes citées influencent la prononciation des différents mots, ils n'ont aucun impact sur leur sens. Dans ce cas, le rappeur n'ignore pas seulement l'orthographe correct du mot comme dans la catégorie I, mais il ne remarque pas l'impact de certains signes sur la phonologie, par exemple la présence d'un « e muet » dans le nom Rockefeller ou la prononciation du double « l » dans le mot « des tranquillisants ». Bien que le rappeur connaisse la prononciation du mot, il n'est pas au courant de son orthographe correcte et il ne prend pas en considération les règles phonétiques. D'autre part, il est nécessaire d'admettre que certaines fautes peuvent être qualifiées de fautes de frappe, comme par exemple « la miséricorde » dans la chanson AKH06 ou « mon côté taciturne » dans la chanson FAY05.

3.2.2.2 Phénomène de l'insertion d'un mot ou d'une phrase

Il ne s'agit que de quatre chansons dont la version « son » diffère de la version « pochette » seulement dans l'insertion d'un mot ou d'une phrase. D'autre part, vingt-deux chansons sont, outre l'insertion d'un mot ou d'une phrase, munies d'une combinaison de plusieurs aspects. D'abord, nous traiterons l'insertion d'un mot en prenant en considération les catégories grammaticales[94]. Nous nous occuperons ensuite de l'insertion d'une phrase dans les versions sonores.

- **Insertion d'un mot**

CGT	ST	P	S	CC
l'interjection	9		ouais	ARS05
			ah ouais	CAN05
			yeah, yeah	KDD05
			ouais	MAC05
			hé	MAF05
			ouais	MCS45
			ouais	PAS17
			yeah	SGL01
			aaao, héyé	SNS01
l'adverbe	6	je m'y attendais pas	je ne m'y attendais pas	ARS05
			non	CAN05
		[...] c'est qui l'poisson rouge ?	[...] c'est qui ce poisson rouge là ?	DLP02
		oui, j'étais funky [...]	oui, j'étais très funky [...]	IAM15
		On va gratter ça [...]	Bon, on va gratter ça [...]	JDE05
		bref, rien de nouveau sous le soleil	bref, rien bien de nouveau sous le soleil	LAR05
le pronom	4	[...] mais l'B.A BA est qu'Léa ou Béa n'aient pas d'aléas	[...] mais l'B.A BA c'est qu'Léa ou Béa n'aient pas d'aléas	MCS65
		Encore en croisade contre l'État avare [...]	Encore en croisade contre c'l'État avare [...]	FFA05

[94] Les différentes catégories grammaticales sont décrites ici : <http://la-conjugaison.nouvelobs.com/regles/grammaire/les-categories-grammaticales-217.php>.

		Mais le ciel peut attendre, j'veux kiffer la vie avant de rendre la mienne	Mais le ciel peut m'attendre, ouais, j'veux kiffer [...]	ARS05
		je me souviens encore [...]	je m'en souviens encore [...]	IAM25
le verbe	4	ceux qui m'aiment me décrivent comme un schizophrène	ceux qui m'aiment me décrivent comme étant un schizophrène	IAM25
		il existé ici	il a existé ici	
		Le petit noir à tête rasée, moi non plus, j'ai pas changé [...]	Le petit noir à tête rasée réapparaît, moi non plus, j'ai pas changé [...]	
			allez	MCS45
le nom	3	Amel, China, Vitaa, Mélissa [...]	Amel, China, Vitaa, Mélissa, Nina [...]	DIA15
		[...] c'est un petit miracle.	[...] c'est un petit miracle, mec.	FFA05
			OK	MAF05
le déterminant	3	la mort est valeur marchande	la mort est une valeur marchande	ASS15
		[...] et condés se tâtent	[...] et les condés se tâtent	IAM25
		[...] c'est que je suis encore sous.	[...] c'est que je suis encore sous un.	DON05
la conjonction de coordination	2	« Hé Disiz, c'est... »	« Hé mais Disiz, c'est... »	DLP02
		Même une goutte sur une vitre...	Mais même une goutte sur une vitre..	JDE05

Tableau n°9 : Phénomène de l'insertion d'un mot

- **Insertion d'une phrase**

CC	**S**
CAN05	[...] j'ai faillit casser ma plume [...]
DLP02	Ce gars-là, c'est qui ?
DLP15	elle se salit
IAM06	Voilà la vérité
IAM15	Ouais, j'étais très funky, allez peace

IAM25	Laisse-moi faire, ami, je travaille pour toi.
KEN25	Petit soldat. – 4x
LSP06	Sur le beat yo.
MIC01	Nanananana, nanananana, Mister You
ORE07	ShiVa – 12x, On y va – 8x

Tableau n°10 : Phénomène de l'insertion d'une phrase

3.2.2.3 Phénomène du changement d'un mot

Le changement de mot concerne vingt-quatre chansons, dont seulement six ne contiennent que le changement d'un mot. Certains changements influencent le sens d'une phrase, d'autres sont sans effet sur l'énoncé. En interprétant ces chansons, le rappeur ne change pas le mot intentionnellement. Les chansons sont classées sous deux catégories et l'influence sur le sens d'une phrase nous sert de critère classifiant.

- **Changement d'un mot sans l'influence sur le sens d'une phrase**

CGT	**ST**	**P**	**S**	**CC**
le déterminant	2	chaque chose	toute chose	FLY05
		[...] y avait la tempête	[...] y avait d'la tempête	IAM25
le pronom	1	La seule vente légale de feuilles de coca. Est pour Coca-Cola qui les raffine pour sa boisson.	[...] qui la raffine pour sa boisson.	ASS15

Tableau n°11 : Changement d'un mot sans l'influence sur le sens d'une phrase

- **Changement d'un mot avec l'influence sur le sens d'une phrase**

CGT	**ST**	**P**	**S**	**CC**
le verbe	8	J'vais sûrement quitter ce monde comme j'ai vécu, seul.	[...] comme je suis venu, seul.	KER15
		Je suis partout.	Je sue de partout.	ARS05
		C'est mon truc.	J'fais mon truc.	CAN05
		tu vois	tu baves[95]	DON05
		certains brillaient	certains priaient	LAL07

[95] Arg. parler, dire (sans idée péjorative), dire des insanités. Centre national de ressources textuelles et lexicales. *Baver : définition de baver* [en ligne]. Page consultée le 2013-07-17. Disponible à l'adresse : <http://www.cnrtl.fr/definition/baver>.

		j'dirais rien	j'ai rien à dire	MAF05
		La Cosa Nostra est son slogan	la Cosa Nostra voilà son slogan	MAF05
		Tu m'fais mal à la caboche, la vie pue d'la gueule, j'vais pas la galoche.	Tu m'fais mal à la caboche, la vie pue d'la gueule, j'prends la galoche.	SNS01
le déterminant	5	C'est mon truc.	J'fais mon truc.	CAN05
		les stéréotypes	ces stéréotypes	CAN05
		l'poisson rouge	ce poisson rouge	DLP02
		cent mille mots à mon service	cent mille mots à son service	IAM25
		mes potes braquent encore, et les miens pas moins...	tes potes...	OXM15
la conjonction	3	Car si les médias ont du pouvoir, il faut savoir pourquoi	Mais si les médias [...]	ASS05
		mais aujourd'hui, je ne suis pas déçu	et aujourd'hui [...]	LSP06
		tu peux être une célébrité et travailler [...]	tu peux être un célébrité, travailler [...]	ORE07
l'adverbe	3	mes bagages sont toujours chargés de pêchés	mes bagages sont encore chargés de pêchés	LAL05
		trop vécue	bien vécue	OXM15
		À quand chaque semaine?	Bientôt chaque semaine?	
l'adjectif cardinal	1	Pense aux quarante-quatre sur la pelouse.	[...] aux cent quarante-quatre[96] sur la pelouse.	MCS45
le pronom X le déterminant	1	à la France d'en haut	à la France de la haut	JDE05
le nom X la catégorie grammaticale inconnue	1	gardave	*le mot inarticulé*	KEN25
la phrase	1	J'ai pas d'sac.	Mais quel sac ?	MAF05

Tableau n°12 : Changement d'un mot avec l'influence sur le sens d'une phrase

[96] Ce changement de l'adjectif cardinal peut être classifié aussi comme insertion d'un mot.

3.2.2.4 Phénomène de l'omission d'un mot ou d'une phrase

Dans l'échantillon examiné, il ne se trouve aucune chanson où l'omission d'un mot ou d'une phrase serait l'unique phénomène touchant le côté textuel. Les onze chansons où un mot ou une phrase ont été omis, contiennent aussi un autre aspect considéré comme une modification du point de vue textuel.

CGT		**ST**	**P**	**S**	**CC**
le pronom personnel		4	On te prend une dent comme trophée d'guerre à la Ramon Dekkers.	On prend une dent [...]	SGL01
			Allez, parlez-moi, j'm'en bats les balloches.	Allez, parlez-moi, m'en bats les balloches.	SNS01
			J'te parle de cette manie [...]	Je parle de cette manie [...]	
			[...] comme Cabrel t'parlait d'la petite Marie	[...] comme Cabrel parlait d'la petite Marie	
l'adverbe		3	Plus rien ne m'étonne.	Plus rien m'étonne.	ORE07
			Comme la violence est gratuite [...]	La violence est gratuite [...]	PAS17
			J'deviens ce que j'n'ai pas voulu être.	J'deviens ce que j'ai pas voulu être.	SNS01
la conjonction		2	Tu peux être une célébrité et travailler [...]	Tu peux être une célébrité, travailler [...]	ORE07
			J'veux écrire dans la souffrance et rapper dans l'aisance.	J'veux écrire dans la souffrance, rapper dans l'aisance	SGL01
le nom		1	Hakeem		KDD05
			Bruits de mitraillette.		PAS17
l'interjection		1	Peace!		IAM15
la préposition		1	Je table sur la qualité, pas sur la quantité [...]	Je table sur la qualité, pas la quantité [...]	IAM25
La combinaison	le nom + l'adjéctif	1	Je dors à poings fermés sous	Je dors à poings	DON05

des catégories grammaticales	qualificatif		anesthésie générale.	fermés sous.	
	le pronom démonstratif + le verbe	1	C'est dans les pires situations [...]	dans les pires situations [...]	OXM15
	le verbe + le nom	1	Ma vie est l'expression d'une volonté dictée avant le big bang	Ma vie une volonté dictée avant le big bang.	ALI05

Tableau n°13 : Phénomème de l'omission d'un mot ou d'une phrase

En considérant le contenu de ce tableau, on peut constater que l'interprète omet avant tout le pronom, l'adverbe ou la conjonction. En général, cette omission n'influence pas radicalement l'énoncé. Par exemple l'interprète de la chanson « J'te parle » fait disparaître le pronom personnel sous forme conjointe, au lieu de « J'te parle de cette manie... », il dit « Je parle de cette manie... ». L'énoncé dans la version « son » n'est pas aussi personnalisé que dans la version « pochette » où il est adressé à l'auditeur, néanmoins son sens reste presque le même.

3.2.2.5 Phénomène du changement de l'ordre des mots

Nous notons le changement de l'ordre des mots auprès de trois chansons dont une (MCS15) ne contient que ce changement. Vous trouverez les exemples précis dans le tableau suivant.

La chanson	**La version « pochette »**	**La version « son »**
MCS15	[...] on nous dit, Dieu est lumière, nous sommes tous frères, mai [...]	[...] on nous dit, nous sommes tous frères, Dieu est lumière, mais [...]
FFA05	Et je veillerai à la famille tant que je vivrai, je l'écrirai.	Et je l'écrirai, je veillerai à la famille tant que je vivrai.
MAC05	[...] en mode crimes, arnaques et gros brolic.	[...] en mode arnaques, crimes et gros brolic.

Tableau n°14 : Phénomème du changement de l'ordre des mots

3.3 Résultats de l'analyse effectuée

Le troisième chapitre de ce mémoire mineur a pour objectif de comparer la version « pochette » et la version « son » des chansons rap. Pour pouvoir réaliser cette comparaison, nous avons choisi deux critères distinctifs, à savoir le côté formel et le côté textuel des différentes chansons.

Le côté formel est conçu comme la description des chansons du point de vue structurel. La structure ordinaire des chansons, soit le couplet 1 – le refrain – le couplet 2, etc., y est souvent élargie d'autres passages. La structure des chansons rap est enrichie de passages additionnels, comme l'intro, l'interlude et l'outro. Il ne s'agit que de 28 % des chansons dont la structure est restée la même dans les deux versions. Si on prend en compte l'occurrence de ces passages ajoutés dans les versions son de l'échantillon analysé, on peut constater que c'est avant tout l'intro et l'outro qui prédomine les autres passages ajoutés. Même s'il s'agit de parties à buts opposés, il faut admettre que ces deux passages partagent quelques traits. Puisque l'intro et l'outro sont introduits « volontairement » dans la chanson au cours de l'interprétation de la chanson, on ne trouve nulle part la définition exacte de leur forme. La plupart des chansons analysées contiennent l'intro ou l'outro sous forme d'exclamation ou de mots scandés. L'interprète scande fréquemment son propre nom ou celui de son collaborateur ou le titre de la chanson. L'intro et l'outro servent souvent à exprimer une dédicace ce qui est un autre trait commun pour ces deux passages. Si on prend en considération les autres types d'intros et d'outros, on s'aperçoit que leurs buts sont complètement différents.

En ce qui concerne les différences entre l'intro et l'outro, il s'agit avant tout de leur orientation. Tandis que l'intro introduit la chanson, l'outro la conclut. De cette raison, l'intro est orientée à attirer l'attention du public, tandis que l'outro résume ce qui a été dit ou éclaire le sens de la chanson en précisant l'énoncé. Il s'ensuit que le texte de l'intro est habituellement court, limité à quelques mots, tandis que le texte de l'outro est plus développé. L'omission de ces deux passages, et avant tout l'omission de l'outro, peuvent donner l'impression que leurs contenus, mentionnés en version « pochette », empêcherait de publier la chanson. De toute façon, nous n'avons pas abordé ce type de cas. Au contraire, il semble qu'une part importante de ces passages

a été ajoutée spontanément à la version « pochette », et tout particulièrement les passages sous forme d'interjection.

En analysant le côté textuel, nous avons classé les divers phénomènes traités en plusieurs groupes, c'est-à-dire les fautes d'orthographe, l'insertion d'un mot ou d'une phrase, le changement d'un mot, l'omission d'un mot ou d'une phrase et le changement d'ordre des mots dans une phrase. Dans l'échantillon analysé, il n'y a que 14 % des chansons, soit 7 chansons, où les deux versions restent les mêmes du point du vue textuel, même si on note la combinaison de plusieurs phénomènes auprès de 58 % des chansons de l'échantillon.

La faute commise la plus souvent est celle d'orthographe. Cependant, le pourcentage n'est pas si élevé si on diminue ce nombre de fautes qui ne sont pas perceptibles dans la version « son » pendant l'interprétation de la chanson. On ne peut les noter qu'en lisant le texte de la version « pochette ». Dans ce chapitre, nous avons mentionné les fautes d'orthographe en les divisant en deux groupes : les fautes d'orthographe ayant une influence sur la prononciation et les fautes d'orthographe sans influence sur celle-ci. En comparant les versions « pochette » et les versions « son », nous ne prenons en considération que les fautes avec l'influence sur la prononciation. Dans la plupart des cas, ces fautes sont commises à cause de l'ignorance de la forme correcte et elles sont liées avant tout à la lettre redoublée, à la ponctuation ou à la conjugaison où les interprètes ne connaissent pas par exemple la forme correcte de l'impératif. Cette ignorance est souvent combinée avec la négligence des règles phonétiques. Bien que le rappeur connaisse la prononciation, il n'est pas au courant de l'orthographe correcte et il ne prend pas en considération ces règles en mettant en page le texte de la chanson. D'autre part, il est nécessaire d'admettre que certaines fautes peuvent être qualifiées de fautes de frappe.

Une autre catégorie traitée du point de vue textuel est l'insertion d'un mot dans la version son au cours de l'interprétation de la chanson. Le chanteur modifie les chansons de cette manière dans vingt-six cas. En général, un mot inséré n'influence pas l'énoncé. En interprétant la chanson, le chanteur insère le plus souvent les mots des catégories grammaticales suivantes : l'adverbe et l'interjection. L'adverbe y sert à souligner l'énoncé, l'interjection s'y réduit à des formes variées pour l'expression

de l'accord, avec la domination de termes familiers comme « ouais » ou l'anglicisme « yeah ».

Au contraire, le sous-chapitre analysant le changement d'un mot en version son comprend beaucoup plus d'exemples de ce phénomène influençant le sens d'une phrase. Cependant, il faut admettre que le changement de sens n'est pas si radical, même si l'énoncé change. Nous supposons que, dans la plupart des cas, l'interprète remplace un terme par un autre inconsciemment. Généralement, la catégorie grammaticale d'un mot remplacé est conservée. La même conclusion est valable aussi pour les versions sonores où le rappeur a omis un mot. Dans plusieurs cas, ce fait n'influence pas du tout le sens de l'énoncé. Il s'agit plus particulièrement du pronom, de l'adverbe ou de la conjonction qui sont oubliés au cours de l'interprétation.

CONCLUSION

En linguistique, un corpus spécialisé est défini comme une base matérielle restreinte en conformité avec l'objectif du corpus. Étant donné que le nôtre est de comparer une version « pochette » et « son » d'un échantillon de chansons rap, notre mémoire mineur est basé sur le corpus d'une cinquantaine de chansons de rap faisant partie du corpus RapCor. La sélection des chansons a été aléatoire, notre corpus n'est pas orienté vers des chanteurs, des albums ou des dates de sortie d'albums concrets. Le corpus est composé de deux versions d'une chanson, de sa version « pochette » et de sa version « son ». Pour pouvoir traiter ces textes, il a fallu ajuster une centaine de documents au total. Après cet aménagement, les textes ont été lancés l'un après l'autre dans le programme *MkAlign* et examiné dans l'onglet ***variation***.

Le programme *MkAlign* offre un éventail de fonctions variées à l'utilisateur. Ses fonctions sont présentées dans le deuxième chapitre, à savoir le chapitre théorique, dont le sujet est la description de ce programme. Dans ces pages, la personne intéressée fait la connaissance du processus de paramétrage de ce programme et de l'enregistrement de textes dans le programme *MkAlign* ce qui est important pour qu'on puisse passer au côté pratique de notre recherche. Pour les collaborateurs adhérant au projet RapCor, nous avons préparé un guide simplifié en tchèque qui se trouve en annexe de ce mémoire. Nous espérons que ce guide contribuera à la diffusion de ce programme entre ces collaborateurs, chargés de l'alignement de versions son et pochette des différentes chansons de rap.

Le côté pratique de ce mémoire mineur repose dans le contenu du troisième chapitre. Ce dernier consiste en comparaison de deux versions d'une chanson de rap à l'aide de l'onglet ***variation*** du programme *MkAlign*. Cet onglet nous permet de visualiser les différences entre la version écrite et la version rappée en un clic de souris. Les chansons choisies ont été analysées de deux points de vue différents, du côté formel et du côté textuel.

Dans ce mémoire, le côté formel est envisagé sous l'angle structurel. En général, une chanson est composée de couplets et d'un refrain. Néanmoins, si on prend en considération la structure la plus développée, il faut compléter cette structure générale d'autres parties, comme une intro, un interlude et une outro. Une chanson peut

contenir tous ces éléments mentionnés et paraître ainsi assez complexe. Notons que la suite des passages se déroule généralement dans l'ordre suivant : l'intro - l'interlude – le refrain – le couplet – l'interlude – le refrain – l'outro.

D'après notre recherche, on peut constater que seulement un quart des chansons reste le même lors de leurs interprétations, par rapport au texte proposé aux auditeurs sur le livret de l'album. Prenant en considération les chansons avec un passage ajouté, on constate que 29 % de ce type de chansons ont été enrichies de l'intro et 42 % de l'outro. La moitié des chansons à plusieurs passages ajoutés a été munie de l'intro et de l'outro en même temps. Ceci fait de l'intro et de l'outro les passages ajoutés le plus fréquemment. On peut déduire de ce fait que la plupart des interprètes préfère ne pas modifier radicalement le corps de leurs chansons, mais ils s'amusent à développer et apporter soit le commencement, soit la fin de leur chanson. Les résultats de notre recherche nous permettent de conclure que le contenu de parties ajoutées n'est pas si scandaleux pour qu'il puisse empêcher la publication de la chanson. Notre première hypothèse n'a pas été confirmée.

Concernant le côté textuel, nous avons examiné des aspects différents comme les fautes d'orthographe, l'insertion, le changement ou l'omission d'un mot ou d'une phrase et le changement d'ordre des mots dans une phrase. Si l'on prend en compte la combinaison de plusieurs aspects, notre analyse nous permet de constater que l'interprète commet le plus souvent une faute d'orthographe. En général, ces fautes n'ont aucune influence sur le sens de l'énoncé et elles sont causées avant tout par l'ignorance de la forme correcte du mot de la part de l'interprète. Au contraire, si on prend en considération des chansons avec une seule aspect changé sur le niveau textuel, on peut dire que les interprètes recourent avant tout au changement de mot. Cet aspect prédomine sur les autres auprès des chansons examinées, même s'il faut admettre que cette domination n'est pas si nette. Notre recherche nous amène à conclure que dans la plupart de cas, les interprètes ne changent pas les mots intentionnellement et qu'un nouveau mot est de la même catégorie grammaticale qu'un mot remplacé. L'influence sur l'énoncé est marginale. Ce fait confirme notre deuxième hypothèse.

Nous pouvons également résumer qu'une chanson de rap typique est composée de trois couplets et d'une intro et d'une outro, ajoutées au cours de l'interprétation de la chanson. Par l'intermédiaire de l'intro, l'interprète s'efforce d'attirer l'attention

d'un auditeur, tandis que l'outro lui sert à conclure ou expliquer l'énoncé d'une chanson. Le rap est souvent défini comme un style de musique aux paroles improvisées ou non et scandées sur un rythme martelé. Il semble que, même si l'improvisation n'est pas un élément obligatoire, les rappeurs ont à cœur de nous démontrer qu'ils sont capables d'improviser librement. Cela peut être aussi la raison pour laquelle on peut noter plusieurs différences entre la version « pochette » et la version « son » dans les chansons de rap plus que dans les autres types de chansons. De ce fait, les chansons de rap représentent un échantillon idéal pour notre analyse.

Pendant notre recherche, nous n'avons fait face qu'à une complication relativement gênante, représentée par l'ajustement des différents textes. La phase de l'unification manuelle a pris beaucoup de temps, parce qu'au moment du téléchargement des textes de la base de données RapCor, la plupart d'entre eux n'était pas munie de la ponctuation précise, ce qui rendait impossible d'enregistrer les chansons directement dans le programme *MkAlign*. D'après nos dernières informations, les chansons enregistrées sur ce stockage devraient être actualisées et leur format unifié, ce qui permettrait de les traiter désormais plus facilement. Ainsi, le corpus examiné pourrait être élargi dans le futur, ce qui palliera à la faible représentativité de notre corpus d'une cinquantaine des chansons. Même si notre échantillon n'a pas de valeur de référence, les résultats de notre analyse ont plutôt pour vocation de servir d'exemple. Nous avons créé un mode d'emploi simplifié pour le programme MkAlign dédié aux étudiants qui souhaiteraient continuer dans cette recherche. Nous croyons que les résultats basés sur l'analyse d'un corpus élargi peuvent aboutir à la rédaction d'une définition précise quant aux différences les plus fréquentes entre les versions « pochette » et « son » des chansons de rap qui pourrait être acceptée en général.

BIBLIOGRAPHIE

Articles et monographie

ČERMÁK, František, « Korpusová lingvistika dnešní doby ». In : ČERMÁK, František et BLATNÁ, Renata (éds), *Korpusová lingvistika: Stav a modelové přístupy*, Praha, Nakladatelství Lidové noviny, 2006, pp. 9 – 18.

DIALLO, David, « La musique rap comme forme de résistane ? » *Revue de recherche en civilisation américaine*, 2009, 1.

FLEURY, Serge, *MkAlign (version 2.0) : Manuel d'utilisation*, Paris : Université Sorbonne Nouvelle Paris 3, 2012, 73 p.

KOCEK, Jan – KOPŘIVOVÁ, Marie – KUČERA, Karel (éds), *Český národní korpus: Úvod a příručka uživatele*, Praha: ÚČNK FF UK, 2000, 156 p. ISBN 80-85899-94-9.

REY-DEBOVE, Josette – REY, Alain (éds), *Le Nouveau Petit Robert de la langue française*, édition 2008. ISBN 978-2-84902-321-1.

RYCHLÝ, Pavel, « Korpusy textů na FI MU ». *Zpravodaj ÚVT MU*, 1997, année VIII, n. 2, pp. 9 - 12. ISSN 1212-0901.

ŠULC, Michal, *Korpusová lingvistika: První vstup*, Praha, Nakladatelství Karolinum, 1999, 94 p.

Sitographie

1.2 Quel type de corpus constituer ? [en ligne]. Page consultée le 2013-10-13. Disponible à l'adresse : <http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2005.ahronian_c&part=90677>.

Centre national de ressources textuelles et lexicales. *Baver : définition de baver* [en ligne]. Page consultée le 2013-07-17. Disponible à l'adresse : <http://www.cnrtl.fr/definition/baver>.

ČERMÁK, František – KOCEK, Jan, *Co je korpus?* [en ligne], Page consultée le 2012-08-24. Disponible à l'adresse : <http://ucnk.ff.cuni.cz/co_je_korpus.php>.

Český národní korpus FF UK. *Korpus SYN* [en ligne]. Page consultée le 2012-08-20. Disponible à l'adresse : <http://ucnk.ff.cuni.cz/syn.php>.

GOUADEC, Daniel, *Outils terminologiques* [en ligne]. Le 2010-10-20 [page consultée le 2012-08-23]. Disponible à l'adresse : <http://www.profession-traducteur.net/outils/outils.htm>.

Encyclopédie Larousse en ligne. *Rap*. [en ligne]. Page consultée le 2013-11-09. Disponible à l'adresse : <http://www.larousse.fr/encyclopedie/divers/rap/85678>.

Le Rap français (Histoire et définition) [en ligne]. Page consultée le 2013-10-15. Disponible à l'adresse : <http://www.rap2banlieue.com/rap-francais/>.

Linternaute.com. *Interlude : définition et synonymes du mot interlude dans le dictionnaire* [en ligne]. Page consultée le 2013-03-02. Disponible à l'adresse : <http://www.linternaute.com/dictionnaire/fr/definition/interlude/>.

Le Nouvel observateur. *La conjugaison* [en ligne]. Page consultée le 2013-06-08. Disponible à l'adresse : <http://la-conjugaison.nouvelobs.com/regles/grammaire/les-categories-grammaticales-217.php>.

MARSHMAN, Elizabeth, *Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie* [en ligne]. 2003 [page consultée le 2013-10-13]. Disponible à l'adresse : <http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf>.

PALA, Karel – RYCHLÝ, Pavel, *Velké textové korpusy v praxi* [en ligne]. Le 2007-07-14 [page consultée le 2012-08-19]. Disponible à l'adresse : <http://www.datakon.cz/datakon08/d07_tut_pala.pdf>.

SYLED [en ligne]. 2009 [page consultée le 2012-08-25]. Disponible à l'adresse : <http://syled.univ-paris3.fr/presentation.html>.

TEUBERT, Wolfgang, La linguistique de corpus : une alternative [version abrégée]. In *Semen*, 27 [en ligne]. Le 2009-04-01 [page consultée le 2012-07-24]. Disponible à l'adresse : <http://semen.revues.org/8914>.

SIGLES

CC	Code de la chanson
CEDISCOR	Centre de recherche sur les discours ordinaires et spécialisés
CGT	Catégorie grammaticale touchée
CR-Trad	Centre de Recherche en Traductologie
ČNK	Český národní korpus (le Corpus national tchèque)
KWIC	Key Word in Context
MVT	Mode de verbe touché
P	version « pochette » de la chanson
S	version « son » de la chanson
ST	« somme totale » des exemples provenant de la même catégorie
SYLED	Systèmes, Linguistiques, Enonciation et Discursivité
TEI	Text Encodint Initiative
TF	Type de la faute
XML	Extensible Markup Language

FIGURES, GRAPHIQUES ET TABLEAUX

FIGURES

GRAPHIQUES

TABLEAUX

ANNEXE 1 : GUIDE SIMPLIFIÉ DU PROGRAMME MKALIGN

I. Stažení a instalace programu

MkAlign 2.00 (2.0b146) pro Windows je k dispozici na tomto odkazu: http://www.tal.univ-paris3.fr/mkAlign/setup-mkAlign.exe (poslední aktualizace z 27. 5. 2013). Pro 64bitovou verzi je pak ke stažení zde: http://www.tal.univ-paris3.fr/mkAlign/setup-mkAlign-x64.exe.

II. Příprava textů

1. Uložení porovnávaných textů ve formátu *.txt (kódování UTF-8)
2. Úprava textů vložením interpunkce

   Abychom nemuseli vkládat žádný specifický dělicí znak, pomocí něhož bude text programem rozdělen, upravíme pouze stávající interpunkci tak, aby byly jasně vyznačeny odstavce.

   V případě programu MkAlign jsou následující znaky považovány za znaky ukončující pasáž: tečka, otazník, vykřičník… Tyto znaky musejí následovat těsně za slovem (tedy bez pevné mezery) a jejich výčet je možné nastavit přímo v programu MkAlign v kartě ***„Param“***. U písní je nezbytné tyto znaky, chybí-li, doplnit a poté vložit znak odstavce ¶ (tlačítko ENTER). Za tímto znakem nesmí následovat žádné skryté znaky (mezera). Ve starších verzích poznámkového bloku nelze skryté znaky zobrazit. Pro usnadnění kontroly je tedy možné text nejprve upravit ve Wordu a následně jej uložit do formátu *.txt.

   Např.

   intro.¶
   Les ombres sont des rêves...¶
   couplet=1.¶
   Cette histoire est une fable, le conteur de celle-ci est fiable (...).[97]

   - u názvů jednotlivých pasáží, tj. intro, couplet 1 atd., smažeme znaky <>. Výsledný dokument, porovnávající verzi pochette a son, je totiž ve formátu html a tyto znaky brání korektnímu zobrazení. Následně doplníme tečku – kdybychom tak neučinili, byl by nadpis programem zahrnut do věty „Cette histoire…“

III. Nahrání textů do programu MkAlign

1. Nastavení programu MkAlign

   Program spustíme pomocí ikony  . V kartě ***„Param“*** upravíme nastavení kódování – „codage source“ a „codage cible“ změníme na UTF-8 (Unicode). V kartě ***„Align“*** zaškrtneme v sekci „Segmenteur“ políčko „prétraitement“ (v sekci „Prétraitement“ ponecháme zaškrtnuté „paragraphe“)

2. Nahrání textů

[97] Akhenaton – « Au fin fond d’une contrée », Métèque et mat, 1995.

V kartě „Align“ klikneme v sekci „Mode général: Chargement source et cible“ na ikonu  a nahrajeme text verze pochette (uložený ve formátu *.txt). Tentýž postup zopakujeme při vkládání verze son přes ikonu  .

IV. Úprava nahraných textů v programu a porovnání obou verzí

1. Zrcadlové srovnání verze pochette a son

   Pokud byly interpretem do verze son přidány některé pasáže (např. intro), je nezbytné srovnat v programu obě verze tak, aby si zrcadlově odpovídaly. Nejčastěji se jedná o **vložení prázdných řádků do verze pochette**. Toho docílíme tak, že klikneme pravým tlačítkem myši na ikonu  (ikona se zaktivní) a následně levým tlačítkem myši na konec řádku (za poslední znak), za něhož si přejeme prázdný řádek vložit. Postup opakujeme tolikrát, kolikrát je nezbytné prázdný řádek přidat.

   V případě, že si naopak přejeme dvě po sobě následující buňky spojit, klikneme pravým tlačítkem myši na ikonu  a poté levým tlačítkem na konec řádku, za něhož chceme obsah následující buňky přesunout. Tento postup můžeme použít i v případě, že omylem vložíme více prázdných buněk, než potřebujeme. Takto se text nepřesune do stejné buňky, ale pouze o buňku výš (tedy do původně prázdné buňky). Text je převeden do stejné buňky, musíme jej však pomocí tlačítka **Backspace** posunout nahoru, aby byl viditelný.

   Mezi jednotlivými stránkami listujeme pomocí šipek umístěných v pravém spodním rohu programu (nejsou-li viditelné, musíme program zvětšit na celou obrazovku). Jsou-li **obě verze zrcadlově shodné**, můžeme přejít k jejich srovnání.

2. Porovnání verze pochette a son

   V programu MkAlign otevřeme kartu ***„Variation“.*** Ponecháme výchozí nastavení a klikneme na ikonu  , která zahájí porovnávání obou nahraných verzí. Vizuálně ověříme, zda si jednotlivé pasáže odpovídají, a následně vzniklé srovnání uložíme. Výsledný dokument je ve formátu html.

ANNEXE 2 : LISTE DES CHANSONS ANALYSÉES

	Les données structurelles	Les métadonnées			
Le sigle	**Le nombre de couplets**	**l'interprète**	**la chanson**	**l'album**	**la date de sortie**
AKH06	3	Akhenaton	Au fin fond d'une contrée...	Métèque et Mat	1995
AKH15	3	Akhenaton	Soldats de fortune	Soldats de fortune	2006
ALI05	2	Ali	Oraison funèbre	Chaos et harmonie	2005
ARS05	3	Arsenik	Chrysanthèmes	Quelques gouttes suffisent	1998
ASS05	3	Assassin	L'éducation à travers les médias	Le future que nous reserve-t-il	1992
ASS15	2	Assassin	Légal ou illégal	L'homicide volontaire	1995
BOO05	3	Booba	La facheuse	Le Panthéon	2005
CAN05	2	Case Nègre	Ma plume	Prologues	2005
DIA07	3	Diams	La boulette	Dans ma bulle	2006
DIA15	3	Diams	Big up	Dans ma bulle	2006
DLP02	3	Disiz La Peste	Le poisson rouge	Le poisson rouge	2000

DLP 15	2	Disiz La Peste	La fille facile	Histoires extraordinaires	2005
DON05	3	Don Choa	Anesthésie générale	Jungle de béton	2007
FAY05	2	Fayçal	Vraies liaisons et lesions	S.O.S.	2009
FFA05	4	Fonky Family	Comme on débarque	Marginale musique	2006
FLY05	2	Flynt	Notre existence	J'éclaire ma ville	2007
IAM06	3	IAM	Le nouveau président	De la planète Mars	1991
IAM15	1	IAM	Crécelle	De la planète Mars	1991
IAM25	2	IAM	Chez le Mac	L'école du micro d'argent	1997
IAM35	3	IAM	Revoir un printemps	Revoir un printemps	2003
JAD05	3	James Deano	Ma vie de célibataire	Secrets de l'oubli	2009
JDE05	2	James Delleck	Gérard de Roubaix	Le cri du papillon	2007
KDD05	3	KDD	Dekadanse	Opte pour le K	1996
KEN03	3	Keny Arkana	J'me barre	Entre ciment et belle étoile	2006
KEN25	4	Keny Arkana	Petit soldat	L'esquisse 2	2011
KER05	2	Kery James	Il y a pas de couleur	Si c'était à refaire	2001
KER15	2	Kery James	Combien	Ma verité	2005

KOM05	2	Koma	Deux pour ta perte	Le réveil	1999
KSH03	3	Kool Shen	Salope.com	Crise de conscience	2009
LAK01	2	La Fouine ft. Kamelancien	Vécu	Capitale du crime Vol 3	2011
LAL05	2	L'Algérino	Entre 2 flammes	Mentalité pirate	2007
LAL07	3	L'Algérino	Allo maman bobo	C'est correct	2011
LAR05	3	La Rumeur	P.O.R.C.	Regain de tension	2004
LSP06	3	Les sages poètes de la rue	Les filles sont belles	Qu'est-ce qui fait marcher les sages	1995
LUN05	2	Lunatic	La lettre	Mauvais Oeil	2000
MAC05	3	Mac Tyer	Mon pote Omar	Le Général	2006
MAF05	2	Mafia Trece	La loi du silence (Suspects usuels)	La Cosa Nostra	1997
MCS15	1	MC Solaar	Concubine de l'hémoglobine	Prose combat	1994
MCS45	3	MC Solaar	Message de l'ange	MC Solaar	1998
MCS65	1	MC Solaar	J'connais mon rôle	Mach 6	2003
MCS75	2	MC Solaar	In God We Trust	Chapitre 7	2007

MIC01	4	Mister You ft. Colonel Reyel	Mets-toi à l'aise	Dans ma grotte	2011
NTM25	3	NTM	On est encore là	Suprême NTM	1998
ORE07	3	Orelsan	Plus rien ne m'étonne	Le chant des sirènes	2011
OXM15	3	Oxmo Puccino	S'13-6.35	L'amour est mort	2001
PAS17	3	Passi	Chambre de gosses	Evolution	2007
ROC05	3	Rocé	Pire que la fiction	Top départ	2001
SEP05	3	VII	Funérailles électriques	Le grand chaos	2009
SGL01	2	Seth Gueko ft. La Fouine	Toucher le ciel	Michto	2011
SNS01	3	Sniper ft. Soprano	J'te parle	À tout épreuve	2011