
MÉMOIRE DE RECHERCHE

Département des Sciences du Langage
M1 Linguistique, Informatique et Technologies du Langage

# DÉTECTION AUTOMATIQUE DE LA DESPÉCIALISATION POUR UN USAGE LEXICOGRAPHIQUE

## LE CAS DU DOMAINE DU SPORT.

Léa DELVENNE
Sous la direction de Cécile FABRE et Franck SAJOUS

2016-2017

# Table des matières

# Index des tables



# Index des figures

# INTRODUCTION

Ce projet de recherche s'inscrit dans une démarche visant à assister le travail lexicographique avec des ressources numérisées et des outils informatiques. Notre objectif est de développer une méthode qui permette de détecter automatiquement les cas de déterminologisation (ou despécialisation) en corpus, dans le but de mettre à jour les articles de dictionnaire des mots concernés.

Nous parlons de déterminologisation lorsqu'un mot initialement employé dans un domaine spécialisé acquiert un emploi plus général. Dans ce mémoire, nous parlons de domaine spécialisé ou de domaine de spécialité de manière indifférenciée pour désigner tous les domaines dans lesquels nous communiquons en employant un vocabulaire spécifique. Les conséquences d'une déterminologisation peuvent être l'apparition d'un nouveau sens dans la langue générale ou la modification du sens d'origine dans la langue de spécialité. Parfois une déterminologisation entraîne ces deux conséquences. Ce phénomène est décrit dans l'article *"L'étirement" du sens terminologique : aperçu du phénomène de la déterminologisation* (Meyer et Mackintosh, 2000).

Le travail de lexicographie réside dans la constitution de ressources dictionnairiques inventoriant l'usage des mots d'une langue le plus fidèlement possible. Repérer automatiquement un phénomène de despécialisation constitue donc une tâche lexicographique qui se justifie pleinement, puisque cela permettrait de surveiller un type particulier d'évolution sémantique afin d'être en mesure d'en rendre compte.

Le développement des outils de traitement automatique des langues (désormais TAL) permet aux linguistes de travailler avec des volumes importants de données. L'avantage est qu'il devient possible de vérifier si des phénomènes observés se répètent de manière significative ou s'il s'agit uniquement de cas isolés. En ce qui concerne la lexicographie, les outils de TAL ont pour objectif de faciliter la constitution et la mise à jour de dictionnaires. Dans ce mémoire, nous nous intéressons principalement à leur mise à jour. Concrètement, cela implique de développer des méthodes de recherche de phénomènes linguistiques témoignant d'une évolution de la langue. Ces évolutions sont diverses, et peuvent se manifester sous forme d'apparition d'un nouveau sens, parfois même d'un nouveau mot (*phablette* fera son entrée dans le *Petit Larousse* 2018 selon *le Figaro*[1]), d'une nouvelle manière d'orthographier un mot (souvenons-nous de la polémique créée autour de la nouvelle manière d'orthographier le mot *o(i)gnon*), de réhabilitation d'un mot auparavant considéré comme vieilli, etc. Si le spécialiste s'aperçoit qu'une évolution se manifeste de plus en plus fréquemment, il devient alors nécessaire d'y porter plus d'attention et, le cas échéant, d'effectuer une mise à jour du dictionnaire.

1 http://www.lefigaro.fr/conjoncture/2017/05/30/20002-20170530ARTFIG00159-gif-phablette-le-numerique-entre-en-fanfare-dans-le-petit-larousse.php

Nous choisissons de réfléchir à la détection de la déterminologisation à l'aide d'un travail sur le contexte. Nous formulons l'hypothèse qu'un mot employé dans une acception spécialisée est entouré d'un contexte constitué d'un vocabulaire propre à ce domaine de spécialité. De fait, s'il est possible de déterminer automatiquement si un contexte est spécialisé, nous pourrons en déduire le type d'emploi étudié. Et si un emploi inhabituel est détecté, c'est-à-dire si un terme spécialisé se retrouve fréquemment employé dans un contexte qui n'a rien à voir avec le domaine de spécialité dont il est issu, il pourra être soumis à un lexicographe pour vérification manuelle. Nous cherchons à déterminer si le contexte est un bon indice pour déduire automatiquement le type d'emploi d'un mot.

Pour mettre en œuvre ce projet, nous travaillons sur le domaine du sport. Nous le considérons comme un domaine de spécialité dans la mesure où il existe un vocabulaire spécifique au domaine sportif et, à la fois, nous supposons qu'il est suffisamment populaire pour que des glissements sémantiques y aient fréquemment lieu.

Nous divisons ce mémoire en cinq chapitres. Tout d'abord nous présenterons notre contexte d'étude. Puis nous détaillerons plus précisément les ressources à notre disposition. Dans un troisième temps nous décrirons la méthode que nous avons choisie, puis nous consacrerons un chapitre à décrire les différentes manipulations effectuées pour tester la validité de cette méthode ainsi que les résultats que nous obtenons. Enfin, nous parlerons des perspectives qu'offrent ces résultats, non seulement pour améliorer la méthode mais également les emplois que l'on peut en faire.

# CHAPITRE I : CONTEXTE D'ÉTUDE

## I. LA DÉTERMINOLOGISATION

### 1. Qu'est-ce que la déterminologisation ?

Décrit d'abord par Guilbert (1975 : 84) puis par Meyer et Mackintosh (2000 : 200), à qui il doit son nom, le phénomène de déterminologisation est relativement peu étudié. Pourtant, ces dernières années, la « société du savoir », qui fait référence à cette société de plus en plus éduquée et qui accorde une grande importance à la connaissance, a contribué à l'amplification du phénomène, d'après Meyer et Mackintosh. Nous parlons de déterminologisation lorsque l'usage de mots appartenant à un domaine bien précis glisse vers d'autres domaines ou, le plus souvent, vers la langue générale.

Pour ce projet de recherche, nous avons fait le choix de nous intéresser à la despécialisation des termes appartenant au domaine du sport. Nous partons des idées de Galisson (1978) qui, dans le premier volume de sa thèse s'intéresse à la diffusion du vocabulaire du football dans la langue générale. Il considère que cette discipline est tellement répandue qu'elle s'inscrit comme élément culturel à part entière et suscite une communication importante, tant parmi les professionnels et les amateurs que parmi le reste de la société (Galisson, 1978 : 13-16). De ce fait, le football donne lieu à de nombreux glissements lexicaux qu'il utilise pour montrer dans quelle mesure le vocabulaire d'une langue de spécialité peut se diffuser dans le langage courant. Nous reprenons cette idée en l'étendant aux autres sports. Le sport paraît en effet être un domaine suffisamment spécifique pour qu'il soit possible d'admettre l'existence d'un vocabulaire qui lui est propre, et également suffisamment populaire pour être un réservoir potentiel de termes en voie de despécialisation.

Prenons l'exemple du mot *marathon* ; la première définition donnée par le TLFi est la suivante :

> **MARATHON**, subst. Masc.
>
> **A.** *SPORTS, JEUX*
>
> **1.** *ATHL.* Épreuve de course à pied de grand fond, sur une distance de 42,195 kilomètres.

*Texte 1: une première définition de* marathon *selon de TLFi.*

Le mot *marathon* désigne en effet une course d'endurance. Il s'agit d'une épreuve sportive fameuse, et le mot est employé à l'origine pour faire référence à cette course ; il appartient bien à un domaine de spécialité, le sport. Toutefois, nous trouvons également des emplois de *marathon* dans d'autres contextes ; en effet, il n'est pas rare de trouver des phrases du type :

> *« Approuvée mardi 21 février après un* ***marathon*** *législatif, la loi, qui fera date, vise à prévenir les violations des droits humains dans les filiales et chez les sous-traitants des entreprises multinationales. » (*Hatchuel Armand (2017)*,* Le devoir de vigilance est une norme de gestion*, Le Monde économie*[2]*)*

Le TLFi recense un emploi différent de ce mot, au figuré, ayant le sens suivant :

> **B.** *Au fig.* Épreuve d'endurance, séance pénible et éprouvante pour les participants en raison d'une durée anormalement longue.
>
> *Texte 2: une définition du sens figuré de* marathon *selon le TLFi.*

Cette définition, ainsi que l'exemple cité plus haut nous montrent bien que le mot *marathon*, initialement propre au domaine sportif, a peu à peu glissé vers des emplois plus généraux. Le mot ne désigne plus forcément la course de 42 kilomètres, mais les notions de pénibilité et de difficulté qu'on imagine inhérentes à cette courses ont été conservées, si bien que *marathon* s'emploie aujourd'hui pour parler d'un moment où l'enchaînement de tâches longues et/ou nombreuses est tel que cela en devient exténuant. Cet exemple est un cas de despécialisation ; un mot employé initialement dans un domaine bien précis (ici, le sport) a glissé vers des emplois plus généraux.

Si, depuis le début de ce mémoire, nous employons indistinctement les mots *déterminologisation* et *despécialisation*, c'est qu'aucun réel consensus n'existe sur la manière de nommer ce phénomène (Humbert-Droz, 2014 : 13). En effet, Guilbert (1975 : 84, dans Humbert-Droz, 2014 : 13) parle de *banalisation* et estime qu'un emploi se banalise parce que les utilisateurs d'une langue de spécialité finissent par se servir de cet emploi dans un contexte plus courant. Le terme est repris par Galisson (1978 : 9), qui le définit comme un processus plus ou moins conscient initié pour faciliter la communication entre les spécialistes et les non-spécialistes. Rastier et Valette (2009), parlent de *dédomanialisation,* phénomène qui *« autonomise un sémème par rapport à son domaine d'origine »* (Rastier & Valette, 2009 : 7, cité par Humbert-Droz, 2014 : 15). Si la manière d'interpréter et de nommer le phénomène diffère, il semble y avoir un certain consensus sur le fait qu'au cours du processus, un mot employé dans une langue de spécialité est récupéré par une majorité de locuteurs, qui le comprennent et s'en servent dans la langue générale. Nous

2 http://www.lemonde.fr/idees/article/2017/03/02/le-devoir-de-vigilance-est-une-norme-de-gestion_5088046_3232.html?xtmc=marathon&xtcr=51

poursuivrons ce travail en continuant de désigner le phénomène étudié par les mots *déterminologisation* et de *despécialisation*.

## 2. Langue générale et langues de spécialité

Afin de mieux comprendre le phénomène de despécialisation, il convient de dégager des pistes de réflexion sur les concepts de *langue générale* et *langues de spécialité*. Dans son mémoire de master sur la despécialisation dans le domaine spatial, Julie Humbert-Droz (2014) s'intéresse à ce qui est entendu par langue générale et langues de spécialité. Elle s'appuie sur les travaux de trois auteurs, Rondeau (1981), Kocourek (1991) et Cabré (1998). Voici ce que nous en retenons :

### a. Langue générale

Rondeau définit la langue générale (qu'il appelle *langue commune*) comme « l'ensemble des mots et expressions qui, dans le contexte où ils sont employés, ne se réfèrent pas à une activité spécialisée » (Rondeau, 1981 : 26 cité dans Humbert-Droz, 2014 : 7). Il est rejoint dans cette idée par Kocourek qui parle de « langue toute entière » (Kocourek, 1991a : 13 cité dans Humbert-Droz, 2014 : 7). Selon Humbert-Droz, ces deux auteurs ont en commun de considérer qu'une partie de la langue est « commune à tous les locuteurs et ne relève pas d'un domaine spécialisé » (Humber-Droz, 2014 : 7). Elle complète la définition en précisant qu'on définit généralement la langue générale comme l'« ensemble de règles, d'unités et de restrictions qui font partie des connaissances de la majorité des locuteurs d'une langue » (Cabré, 1998. : 115 cité dans Humbert-Droz, 2014 : 8).

Pour résumer, la langue générale englobe tous les mots dont l'emploi est connu et maîtrisé par une majorité de locuteurs. Une langue que nous utilisons pour communiquer et nous faire comprendre des autres locuteurs au quotidien, et qui ne requiert pas l'utilisation systématique d'un vocabulaire spécifique.

### b. Langues de spécialité

Selon Rondeau, les langues de spécialité désignent *« une communication liée à une activité spécialisée et [il] parle alors de termes et non plus de mots ou d'expressions »* (Humbert-Droz, 2014 : 8, se référant à Rondeau, 1981 : 26). À cette idée, vient s'ajouter celle de Kocourek selon laquelle une langue de spécialité serait une *« variété de la langue entière »* (Kocourek, 1991a : 13, cité dans Humbert-Droz, 2014 : 8). Cette notion de variété de la langue entière est à rapprocher de celle de « sous-langage » développée par Harris, qui considère que le langage peut se découper en plusieurs sous-langages avec des propriétés grammaticales qui leur sont propres et que le « langage dans son ensemble » ne possède pas (Harris, 1968 : 170-171). Il faut en outre préciser

que Rondeau et Cabré parlent tous les deux des frontières poreuses entre la langue générale et la langue de spécialité, ce qui explique selon eux pourquoi certains termes de spécialité se retrouvent dans la langue générale et vice versa.

Après avoir étudié les points de vue de ces auteurs, nous pouvons donc dire que les langues de spécialité sont en quelque sorte des sous-parties de la langue « entière », régies par un vocabulaire et des codes de communication qui leur sont propres. Les frontières entre langues de spécialité et langue générale sont poreuses, ce qui permet des glissements terminologiques dus aux évolutions de la langue dans le temps, donc des déterminologisations.

Dans la suite de ce mémoire, le mot *terme* se réfère, comme dans la littérature, aux mots employés dans un domaine de spécialité.

## 3. Typologie des déterminologisations

Meyer et Mackintosh distinguent plusieurs formes de déterminologisation. Pour commencer, elles différencient les cas où le sens est conservé des cas ou le sens évolue.

### a. Lorsque le sens ne change pas

Dans les cas où le sens est conservé, nous observons deux phénomènes. Le premier se traduit par une réappropriation de l'emploi de spécialité par une majorité de locuteurs. Rien ne change au niveau du sens ou de la morphologie du terme, il est simplement employé par un plus grand nombre de personnes et n'est plus cantonné au seul domaine dont il est issu. Les auteures illustrent ce cas par l'exemple de *VIH.* Initialement utilisé en médecine, ce terme est devenu beaucoup plus fréquent au moment ou l'épidémie a pris de l'ampleur ; son emploi s'est généralisé sans pour autant entraîner une modification du sens initial. Dans le sport, il est difficile d'observer ce genre de cas parce qu'il s'agit d'un domaine déjà assez populaire, présent dans le vocabulaire de beaucoup de locuteurs. Là où cela semble flagrant pour *VIH* qui est passé du domaine médical à la langue générale en peu de temps et sans changer de sens, c'est plus compliqué pour un terme tel que *football,* par exemple : ce mot a toujours fait référence au jeu de ballon et il est très utilisé dans le langage général. Pouvons-nous pour autant dire qu'il a subi une despécialisation ? Ce type de despécialisation demande une réflexion sur la fréquence d'apparition d'un terme dans un contexte différent de celui où il apparaît habituellement et sur le seuil de fréquence à partir duquel on le considère comme déterminologisé. En outre, comme aucun changement de sens n'a lieu, et que la différence mise en avant par Meyer et Mackintosh se situe surtout au niveau de la fréquence d'emploi, nous ne pouvons pas dire que cela entraîne une obsolescence de l'article concerné dans le dictionnaire et un besoin de mise à jour. VIH s'est popularisé, mais il appartient toujours au domaine médical au même titre que football appartient au domaine sportif. Nous

n'excluons donc pas ce type de déterminologisation dans ce mémoire, mais notre priorité reste de détecter les cas qui nécessitent une mise à jour du dictionnaire.

Le deuxième cas de figure observé consiste à se réapproprier le sens d'un mot, mais en changeant de terme pour le désigner. Ici, les auteures illustrent leur propos par l'exemple du terme *encéphalopathie spongiforme*, qui, en se despécialisant, a donné *maladie de la vache folle*. Ici encore, peu d'exemples disponibles dans le domaine sportif qui est en général suffisamment compréhensible par tous pour que les locuteurs ne ressentent pas le besoin de changer les mots. Lorsqu'un mot est changé, il s'agit plutôt d'un acte de vulgarisation, comme pour le terme *bombe*, qui désigne la protection pour la tête utilisée en équitation et qui est souvent appelée casque par les personnes qui ne connaissent pas bien l'équitation. Le processus est différent car un mot ou une expression n'est pas vraiment créée, il s'agit simplement d'utiliser un hyperonyme existant pour parler de cet objet.

### b. Lorsque le sens évolue

Le cas suivant se manifeste différemment : le mot ne change pas, mais le sens évolue quelque peu. Meyer et Mackintosh écrivent que bien souvent, nous assistons à une « dilution » (Meyer et Mackintosh, 2000 : 204) du sens initial, c'est-à-dire que certaines notions véhiculées par l'emploi initial sont conservées, mais le sens n'est plus aussi précis, il est employé au figuré. Elles donnent l'exemple du mot *recycler*, qui faisait initialement référence au processus écologique visant à utiliser les matières premières des déchets pour en faire de nouveaux objets. Rapidement, *recycler* est employé dans le milieu professionnel pour parler des nouvelles formations auxquelles sont soumis les professionnels pour répondre à des exigences nouvelles. Le sport regorge de cas semblables. Souvent, il s'agit d'expressions qui sont réemployées au quotidien. En voici quelques exemples :

L'expression *« faire le grand écart »* désigne initialement une pratique en danse ou en gymnastique qui consiste à écarter les jambes jusqu'à toucher le sol et former un angle de 180 degrés. La difficulté de l'exercice a donné lieu à une réappropriation de l'expression dans des emplois plus généraux faisant référence à la difficulté de concilier des activités, des relations, des idées très différentes voire opposées.

*« Être dans les starting-blocks »*. En athlétisme, les starting-blocks sont les modules que l'on place sous les pieds des coureurs pour les caler au moment du départ. Initialement, cette expression signifie donc « être sur le point de démarrer sa course ». Nous trouvons également d'autres emplois dans des contextes différents, et qui signifient être sur le point (voire impatient) de commencer quelque chose.

Des mots comme champion, marathon ou relais possèdent des emplois déterminologisés ; nous pouvons les retrouver dans le contexte général avec un sens qui a légèrement évolué mais

qui conserve des liens avec le sens initial.

Dans ce mémoire, nous nous concentrons surtout sur le dernier type de déterminologisation évoqué, car c'est celui qui est le plus susceptible d'entraîner une mise à jour du dictionnaire. Nous ne sommes pas en mesure de détecter des déterminologisations qui entraînent un changement de mot. En effet, la méthode envisagée, qui sera expliquée dans les grandes lignes à la fin de ce chapitre puis plus en détails au chapitre III (p. 29), ne prévoit pas de détecter, pour un emploi d'un mot-cible, tous les autres mots dont le sens correspond à l'emploi despécialisé de ce mot-cible. Il s'agit plutôt d'analyser le contexte d'un mot-cible pour savoir si ce contexte paraît spécialisé ou non, et d'en déduire le type d'emploi du mot-cible. Cela ne fonctionne que si l'emploi despécialisé s'exprime avec le même mot que l'emploi spécialisé. En revanche nous n'excluons pas de trouver un cas où le sens ainsi que le terme n'ont pas évolué, mais se sont simplement popularisés, ce qui correspond au premier type de déterminologisation décrit.

# II. DÉTERMINOLOGISATION ET LEXICOGRAPHIE

Dans cette partie nous expliquons certains éléments qui permettront de mieux comprendre le fonctionnement de la lexicographie aujourd'hui, et nous revenons sur le lien entre la déterminologisation et la lexicographie.

## 1. La lexicographie informatisée : assister le travail du lexicographe grâce au TAL

La lexicographie est la discipline qui s'attache à recenser, décrire, classer et illustrer les différents mots et emplois d'une langue, afin de les compiler dans un dictionnaire. Aujourd'hui ce travail est en grande partie informatisé, et le rôle du TAL est de mettre au point des outils pour automatiser, au moins partiellement, le travail du lexicographe. L'informatisation et l'automatisation des tâches lexicographiques présentent un certain nombre d'avantages, dont celui de pouvoir traiter un volume de données beaucoup plus important que si tout était manuel.

### a. Le travail lexicographique

De sa conception à son édition, et même une fois édité, diverses tâches jalonnent la fabrication d'un dictionnaire. Une fois le type de dictionnaire défini et les questions de mise en page réglées, l'aspect linguistique du travail devient central. Il s'agit de parvenir à décrire clairement le ou les sens et usages d'un mot de la manière la plus générale, la plus représentative et la plus compréhensible possible. Une entrée de dictionnaire peut comporter des informations

telles que la prononciation, les variantes orthographiques, des indication sur la fréquence (courant, rare,...), la ou les formes fléchies, l'étymologie, la ou les définitions, un ou plusieurs exemples pour illustrer la définition, éventuellement des traductions dans d'autres langues et des marques d'usage qui permettent de savoir facilement dans quel contexte s'emploie le mot décrit. Il est également possible d'y trouver des précisions sur la nature grammaticale du mot et sur ses constructions les plus courantes (Atkins et Rundell, 2008, 204-234). L'enjeu du lexicographe est de renseigner ces informations de la manière la plus exhaustive possible, en tout cas pour les éditions monolingues générales destinées aux adultes.

**CHAMPION**, ONNE, subst.

**A.** *Subst. masc., HIST.* Celui qui combattait en champ clos pour défendre la cause d'une autre personne ou la sienne propre. *Un brave, un vaillant champion; il s'offrit à cette dame pour être son champion. Ces champions sont des braves, dont le métier est de se battre pour le compte d'autrui, en combat judiciaire, à l'épée ou au bâton* (FARAL, *La Vie quotidienne au temps de st Louis,* 1942, p. 77).

*P. ext.* Combattant de grand mérite. *Fam.* Tout combattant opposé à un autre. *Un grand vieillard (...) qui avait dû, au temps des guerres de Vendée, être un rude champion, un redoutable adversaire des Bleus* (PONSON DU TERRAIL, *Rocambole,* t. 1, *L'Héritage mystérieux,* 1859, p. 403).
[...]

**B.** *Subst. masc.* ou *fém.*

**1.** *SP.* Athlète ou équipe qui a remporté la première place dans un concours, une épreuve sportive ou un match organisé pour l'obtention de ce titre. *Elle était championne internationale de golf* (S. DE BEAUVOIR, *Mémoires d'une jeune fille rangée,* 1958, p. 152).

*P. ext.* Athlète de premier ordre représentant son pays (*cf. athlète* ex. 3) :

> 3. Je n'ai qu'un tout petit rôle vous savez. Je suis **champion** *du monde* de boxe et le jeune premier qui n'est autre que Valmègue me met nokaoute au troisième round et devient à son tour **champion** *du monde.*
> QUENEAU, *Loin de Rueil,* 1944, p. 173.

**SYNT.** *Un très grand champion; un champion olympique; champion de France, du monde; champion de boxe, de ski; des titres de champion.*

*JEUX.* *Champion de bridge, d'échecs.*

**2.** *Fam.* Personne qui excelle dans une activité, un domaine quelconque. *La mère Bertine, championne du battoir et du pilon à beurre* (H. BAZIN, *Vipère au poing,* 1948, p. 43; *cf.* aussi *athlète* ex. 5).

*Emploi adj., fam.* ou *pop.* De premier ordre, excellent. *Cette fille est championne!; pour faire des gaffes, il est champion.*

*Adj. inv.* [En parlant de pers., d'actes] Digne d'admiration. *C'est champion, c'est un coup champion* (ROB. *Suppl.* 1970). *Alors on voyage aussi? Champion, hein, l'Acropole?* (Daninos ds COLIN 1971) :
[...]

**Rem.** Le fém. est inus. sauf en des emplois fam. au sens A. En revanche, il est empl. dans le vocab. des sp. au XX$^{e}$ s. sans nuance péjorative.

**Prononc. et Orth. :** [ʃɑ̃pjɔ̃], fém. [-pjɔn]. Ds *Ac.* 1694-1932. **Étymol. et Hist. I.** *Champion* **1.** *ca* 1100 *campiun* « celui qui combat en champ clos pour soutenir une cause » *(Roland,* éd. J. Bédier, 2244); 1130-60 *champïon* (*Couronnement de Louis,* éd. E. Langlois, 501 ds T.-L.) - av. 1507 (MOLINET, *Chron.,* éd. G. Doutrepont et O. Jodogne, t. II, p. 313 [année 1492]), devenu ensuite terme hist.; 1704 [en Angleterre] *champion du roi (Trév.)*; **2.** p. ext. 1552 « combattant quelconque (pour une cause) » (EST.); 1668 « homme qui combat » (LA FONTAINE, *Fables,* I, 13 ds LITTRÉ); n'est plus empl. dans la lang. class. que dans le style burlesque (BRUNOT t. 4, p. 329); **3.** 1560 fig. « celui qui défend une pers., une cause » (E. PASQUIER, *Recherches de la France,* 799 ds *IGLF*); [...]

*Texte 3 : Extraits de l'article du mot champion dans le TLFi*

L'article ci-dessus (texte 3) provient du TLFi[3], il s'agit de l'article décrivant les sens et usages du mots *champion*. Nous avons surligné les différents éléments cités précédemment. En rouge, le titre de l'article et le mot dont il est question. En rose clair à côté, sa forme fléchie, en bleu, des indications grammaticales et d'usage sur le mots et ses différents emplois, en vert, les marques lexicographiques, en rose foncé, les définitions, en jaune les exemples, en violet les syntagmes, en orange, les sections prononciation, orthographe, étymologie et histoire. Nous visualisons bien ici les différents éléments d'un article et la façon dont ils s'organisent.

## b. Les ressources

Pour décrire le sens des mots, les lexicographes s'appuient en partie sur leur intuition, qui leur permet d'identifier certains phénomènes ou certains emplois. Mais l'intuition ne suffit pas à faire du dictionnaire une ressource fiable, car la connaissance d'une langue par un locuteur a ses limites. Depuis les années 1960, la lexicographie anglophone est informatisée, et le corpus est la ressource principale pour le travail lexicographique depuis les années 1980 (Atkins et Rundell, 2008 : 3 & 53). En France, le corpus électronique est utilisé pour la première fois en lexicographie lors de la création du TLF dans les années 1970 (Pruvost, 2000 : 55).

Sinclair définit le corpus comme *« une collection d'extraits de langage écrit, au format électronique, sélectionnés en fonction de critères objectifs et aussi représentatifs que possible d'une langue et de ses variétés pour servir de source de données pour la recherche linguistique »* (Sinclair, 2005 : 16, cité par Atkins et Rundell, 2008 : 54). L'intérêt principal du corpus est que les mots apparaissent dans leur contexte, ce qui permet de savoir quel type d'emploi existe et quelle est sa fréquence. Cela permet au lexicographe de décider quoi insérer dans le dictionnaire. Il est cependant important de préciser qu'aucun corpus de saurait être parfaitement représentatif de la langue dans sa globalité, car aucun corpus ne peut contenir tous les usages d'une langue.

Un corpus représente généralement un volume important de données linguistiques, et l'exploiter demande des outils que le TAL est en mesure de fournir à la lexicographie. Il s'agit de concevoir des moyens de détecter automatiquement les informations dont les lexicographes souhaitent rendre compte dans leur dictionnaire. Cela peut se résumer à chercher l'existence d'un mot et sa fréquence pour savoir s'il y a lieu de lui créer une entrée, mais il existe également des outils qui trouvent des collocations et autres informations sur les usages d'un mot. Un concordancier KWIC (keyword in context), par exemple, est un outil basique de la lexicographie. Il permet de chercher dans le corpus et de manipuler des mots en fonction de leur lemme (forme non fléchie), leur classe grammaticale et du type de document dans lequel ils apparaissent (Atkins et Rundell, 2008 : 104-105). Ces outils facilitent l'analyse de corpus et accélèrent le travail linguistique, ce qui permet aux lexicographes de rendre compte de l'usage réel des mots, d'être plus productifs et de se concentrer sur d'autres tâches.

3 http://stella.atilf.fr/Dendien/scripts/tlfiv5/saveregass.exe?303;s=1270144035;r=9;;

## 2. Comment la déterminologisation peut-elle faire évoluer une entrée de dictionnaire ?

Une fois le dictionnaire créé, il doit rester représentatif de la langue dont il dresse l'inventaire. C'est pourquoi il doit être mis à jour lorsqu'une évolution de la langue est constatée et considérée comme suffisamment nette pour être recensée. La plupart des éléments d'une entrée de dictionnaire peuvent évoluer. Et il peut être nécessaire de rajouter une entrée lorsqu'un nouveau mot apparaît.

Comme décrit dans la partie précédente (cf. partie I. 3. p. 9), la déterminologisation peut avoir plusieurs conséquences sur la langue. Lorsque le sens initial du mot dans la langue de spécialité a subi une évolution une fois le mot employé dans la langue générale, la définition peut devenir obsolète. Cela entraîne la nécessité d'ajouter une définition en précisant le type d'emploi et le contexte dans lequel est utilisé cet emploi. Par exemple, le mot *champion*, dont la définition du TLFi se trouve p.12, est utilisé en sport depuis 1877, mais fait à présent partie du langage courant et désigne toute personne douée dans une activité quelconque. À ce titre une nouvelle définition a été ajoutée à l'article pour recenser cet emploi. Lorsque la déterminologisation n'entraîne pas un nouveau sens, il peut tout de même être nécessaire de préciser dans le dictionnaire que le mot dont il est question a également un autre type d'emploi. Le texte 4 est un extrait de l'article correspondant au mot *relais* dans le TLFi. Il y est précisé que le mot relais employé en athlétisme est également utilisé au figuré dans le langage courant.

**RELAIS**[2], subst. Masc.
**[...]**
**B.** [À propos de pers.]
[…]
**b)** *ATHL.* Formule de courses se disputant sur un parcours divisé en quatre sections égales où s'affronte, chacun à son tour, chaque représentant des diverses équipes en compétition (d'apr. PETIOT 1982). *Courses de relais; un relais quatre fois cent mètres. Un bon relais, c'est celui où le bâton se prend entre douze et quinze mètres* (R. BOISSET, *À vos Marques,* 1949 ds PETIOT 1982):

*Les coureurs de* ***relais***
*Tous quatre lancés comme une seule arme, comme une seule bête, comme une seule barque,*
*le plus grand à la poupe et le plus petit qui est en avant,*
*et moi engrené au milieu, moi organe de ce corps vivant,*
*et tous portant les mêmes couleurs, et tous marqués de la même marque...*
MONTHERL., *Olymp.,* 1924, p. 331.

*Au fig. Prendre le relais.* Succéder à quelque chose ou quelqu'un dans la poursuite d'une opération, d'un processus. *Dans ce pays jeune* [*la Russie*], *sans tradition philosophique, de très jeunes gens (...) un « prolétariat de bacheliers » a pris alors le relais du grand mouvement d'émancipation de l'homme, pour lui donner son visage le plus convulsé* (CAMUS, *Homme rév.,* 1951, p. 187). *Chercher une source de profit qui* prenne le relais *du gisement du gaz naturel de L.* (*L'Express,* 3 nov. 1969 ds GILB. 1980).

*Texte 4 : extrait de l'article du mot relais dans le TLFi.*

Enfin, si la déterminologisation entraîne un changement de mot, comme *encéphalopathie spongiforme* qui a donné *maladie de la vache folle*, cela peut donner lieu à la création d'une nouvelle entrée, même si celle-ci renvoie simplement au mot initial, ou bien à une précision dans l'entrée initiale que la version familière de *encéphalopathie spongiforme* est *maladie de la vache folle*.

Ainsi le travail sur la détection automatique de la despécialisation trouve-t-il une justification dans le domaine lexicographique. Nous nous attachons maintenant à expliquer comment utiliser le dictionnaire et les principes de la lexicographie de corpus pour détecter des termes déterminologisés.

# III. REPÉRER AUTOMATIQUEMENT LA DÉTERMINOLOGISATION

Nous décrivons ici les prérequis à la détection automatique de termes déterminologisés.

## 1. Définir des mots-cibles

Chercher à repérer une despécialisation implique en premier lieu de définir des mots-cibles à surveiller en contexte. La despécialisation touche certains mots dont l'emploi est initialement propre à un domaine de spécialité, c'est donc ces mots-là qu'il convient de surveiller. Pour définir les mots à surveiller, nous nous aidons du dictionnaire.

Dans un dictionnaire, pour indiquer qu'un terme a un emploi spécifique ou une particularité quelconque, il existe un élément de la micro-structure appelé marque lexicographique (Atkins et Rundell, 2008 : 226). Une marque lexicographique peut servir à indiquer la fréquence d'usage d'un mot (RARE, TRÈS RARE, COURANT,...), son origine (FRANCE, ANGLETERRE, AFRIQUE, LORRAINE,...), son niveau discursif (FAMILIER, ARGOT, LITTÉRAIRE,...), etc. Pour repérer les termes employés dans une langue de spécialité bien précise, il faut s'intéresser aux marques de domaine. Le texte 5 est un extrait de l'article consacré au mot *marathon* dans le TLFi. Nous y observons qu'il est découpé une première fois en deux catégories : la première est notée « *SPORTS, JEUX* » tandis que la seconde est notée « *au fig*. ». Dans la première catégorie, nous lisons également « *ATHL* » et « *NATAT.* ». Ces indications en lettres capitales sont des noms de domaine. Ils servent à indiquer à l'utilisateur du dictionnaire que le mot *marathon* est un terme relevant du sport, et plus précisément aux catégories sportives *ATHLÉTISME* et *NATATION*.

**MARATHON**, subst. Masc.
**A.** *SPORTS, JEUX*
***1.****ATHL.* Épreuve de course à pied de grand fond, sur une distance de 42,195 kilomètres. *Un coureur de marathon*; *le vainqueur du marathon*. *Les courses de fond comprennent les épreuves de fond court, 5 000 et 10 000 mètres, et de grand fond allant jusqu'au marathon* (R. VUILLEMIN, *Éduc. phys.*, 1941, p. 134).
**2.** *P. anal.*
**a)** *NATAT.* Épreuve prolongée exigeant une grande résistance. *Le crawl (...) est également nagé avec succès dans les marathons, telle la traversée de la Manche que fit, en 1926, la jeune Américaine Ederle* (*Jeux et sports*, 1967, p. 1567).
**b)** [En appos. ou comme 2e élém. de subst. composé] *G. (champion de tennis) semble très éprouvé par son match-marathon de mardi et mercredi contre P.* (*Le Monde*, 29 juin 1969 ds GILB. 1971). *Durant toute la semaine, ils se sont livrés à un match marathon. Douze parties par jour, pour le plaisir et pour l'honneur: les championnats de billard ne comportent aucun prix en espèces* (*L'Express,* 25 mars 1974 ds GILB. *Mots contemp.* 1980).
**3.** *Marathon de la danse. Dans les années vingt et trente, ces «Marathons de la Danse» d'un genre spécial, qui duraient des semaines entières, et au cours desquels des dizaines de couples devaient danser, ou marcher, sans autre répit que des pauses-express de dix minutes toutes les heures* (*Cinéma 70*, 1970, no 149, p. 50).
**B.***Au fig.* Épreuve d'endurance, séance pénible et éprouvante pour les participants en raison d'une durée anormalement longue. *Cet échange de lettres demandait un farouche effort humain. Cela devenait un perpétuel marathon* (LA VARENDE, *Tourville*, 1943, p. 141). *Wahid organise une soirée costumée orientale. Nous jouons. Mlle Debar m'étonne dans le rôle du Sphinx. C'est un marathon. Elle s'y crève* (COCTEAU, *Maalesh*, 1949, p. 86).
[Souvent dans un cont. écon., pol.] *L'accord a été finalement conclu dans le décor habituel des marathons diplomatiques bruxellois* (P. FABRA ds *Le Monde*, 12 mai 1966):
Rapidement, les Six, galvanisés, manifestent une volonté farouche d'aboutir, au prix du **marathon** le plus épuisant que l'Europe ait jamais connu.

*L'Express*, 29 mars 1971, p. 20, col. 3. […]

*Texte 5: extrait de l'article marathon du dictionnaire TLFi.*

Nous décidons donc d'utiliser ces marques de domaine pour établir une liste de mots-cibles à surveiller. Nous introduisons dans le chapitre II le dictionnaire utilisé (cf. chapitre II, partie I. p.21 ) et dans le chapitre III, la méthode utilisée pour constituer la liste de mots-cibles (cf. chapitre III, partie II. 1. p.33).

## 2. Le lien entre contexte et emploi

Dans son ouvrage *Trust the Text, language, corpus and discourse,* Sinclair (2004) développe l'idée selon laquelle les mots sont des unités lexicales qui entrent en relation avec d'autres mots pour former du sens (Sinclair, 2004 : 25). Cette idée a été initialement développée dans les *English Collocation Studies* (2004), qu'il mène dans les années 1970 en collaboration avec Susan Jones et

Robert Daley. Nous en déduisons que le plus sûr moyen de distinguer en corpus un emploi spécialisé d'un emploi général est de se servir du contexte dans lequel l'emploi apparaît. Les deux phrases suivantes sont extraites d'articles de presses des journaux *l'Equipe* et *Le Monde économie*. Elles constituent deux emplois différents du mot *marathon*.

> *« Eliud Kipchoge a décidé de faire l'impasse sur les **Championnats du monde** d'**athlétisme** à Londres (4-13 août). Le Kényan a indiqué avoir encore besoin de repos après sa tentative de **courir** le **marathon** en moins de deux heures en avril. »*
> (Mounic Alain (2017), Eliud Kipchoge renonce au marathon des Mondiaux, *L'Equipe*[4])

> *« Approuvée mardi 21 février après un **marathon** législatif, la loi, qui fera date, vise à prévenir les violations des droits humains dans les filiales et chez les sous-traitants des entreprises multinationales. » (*Hatchuel Armand (2017), Le devoir de vigilance est une norme de gestion*, Le Monde économie*[5]*)*

Le premier emploi est un emploi relevant du domaine sportif. Nous trouvons dans son contexte des mots qui permettent d'identifier cet emploi (en gras). En revanche, dans le second exemple, marathon n'est pas un emploi relevant du sport. Or, il n'apparaît pas dans un contexte sportif. Outre les idées de Sinclair, nous avons mentionné dans la partie I. 2. b. de ce chapitre (cf. p.8) qu'une langue de spécialité permettait une communication liée à une activité spécialisée, et donc se composait d'un vocabulaire propre à cette langue de spécialité. Cela permet d'expliquer que, dans l'exemple ci-dessus, l'emploi spécialisé se trouve dans un contexte spécialisé. Il s'agit en effet d'un acte de communication s'adressant au amateurs de sport et donc le vocabulaire est adapté aux locuteurs. Si nous nous référons aux idées de Kocourek sur la langues spécialisées ainsi que sur les théories de Sinclair, Jones et Daley, nous pouvons formuler l'hypothèse qu'il est possible de détecter automatiquement un type d'emploi (spécialisé ou non) en analysant le contexte dans lequel il apparaît.

## 3. Méthode générale envisagée

Nous décrivons ici comment nous envisageons de détecter les cas de despécialisation des termes sportifs pour un usage lexicographique à partir de l'hypothèse que nous venons d'énoncer. Cette méthode ne pourra être mise en œuvre que si l'hypothèse est vérifiée, c'est ce que nous nous attachons à faire dans la suite de ce mémoire. La figure 1 illustre la méthode que nous nous apprêtons à décrire.

4 https://www.lequipe.fr/Athletisme/Actualites/Eliud-kipchoge-renonce-au-marathon/803987

5 http://www.lemonde.fr/idees/article/2017/03/02/le-devoir-de-vigilance-est-une-norme-de-gestion_5088046_3232.html?xtmc=marathon&xtcr=51

Dans un premier temps, une liste de mots-cibles à surveiller est définie à partir du dictionnaire sur lequel le lexicographe travaille. Le dictionnaire constitue la ressource que nous souhaitons tenir à jour, d'où l'intérêt de surveiller directement les termes qu'il contient. La liste doit être composée de mots qui répondent à des critères objectifs. Il semble plus simple dans un premier temps de ne s'intéresser qu'à des mots monosémiques, dans la mesure où ils sont supposés n'apparaître que dans leur acception sportive, donc dans des contextes liés au sport. De ce fait, un emploi apparaissant dans un contexte non sportif sera automatiquement considéré comme inhabituel, et mis de côté. En revanche, les mots polysémiques peuvent aussi faire l'objet d'une despécialisation, mais comptent déjà différents emplois et sont susceptibles d'apparaître dans des contextes variés. Ces différents emplois doivent être pris en compte lors de l'analyse, il s'agit donc d'un traitement plus élaboré. Un moyen de les intégrer devra être envisagé par la suite. L'objectif de cette étape est d'automatiser la sélection des termes, pour que le lexicographe n'ait pas à s'en occuper.

Une fois la liste de termes-cibles définie, nous observons les occurrences des mots de cette liste en contexte afin de déterminer si l'occurrence apparaît dans un contexte spécialisé ou non. Pour cela, nous avons besoin d'un corpus de textes : comme précisé précédemment (voir partie II. de ce chapitre, p. 11), le lexicographe se sert du corpus comme support de réflexion, pour vérifier ses intuitions et justifier les choix qu'il effectue. Il y analyse des extraits de langue qui lui permettent de savoir ce qui se dit ou non. Le rôle du corpus dans cette méthode est de nous fournir un ensemble de contextes concrets et représentatifs de la langue qui serviront de base à l'analyse automatique. Il s'agit en quelque sorte d'effectuer le travail de recherche et de détection à la place du lexicographe pour qu'il n'ait plus qu'à interpréter les résultats. Un programme analyse le corpus et, pour chaque lemme, regarde s'il fait partie de la liste de mots-cible. Si non, il continue, si oui, le contexte de l'occurrence du mot-cible est analysé et chaque mot du contexte est comparé au contenu d'un lexique spécialisé. Ce lexique spécialisé est constitué de termes appartenant au domaine du sport. Il contient des mots du dictionnaire, mais également d'autres mots susceptibles de participer à la création d'un contexte sportif. Il doit être approvisionné automatiquement et régulièrement pour rester représentatif de la langue de spécialité, même si elle évolue.

Une fois l'ensemble du contexte comparé au lexique, la proportion de mots du contexte qui font partie du lexique est calculée et on attribue à l'occurrence un score qui évalue son degré de spécialisation. En fonction de ce score, l'occurrence est considérée comme spécialisée ou non, et si ce n'est pas le cas, elle est consignée dans un fichier de sortie. À partir d'une certaine proportion d'occurrences despécialisées, le lexicographe est alerté et peut vérifier s'il s'agit ou non d'une déterminologisation et si celle-ci est susceptible d'entraîner une modification du dictionnaire. L'apport de cette méthode réside dans le fait que le lexicographe n'a plus besoin de regarder manuellement tous les mots, pour savoir s'il faut mettre à jour l'article correspondant. Il n'a à se concentrer que sur ceux dont l'emploi est automatiquement considéré comme inhabituel.

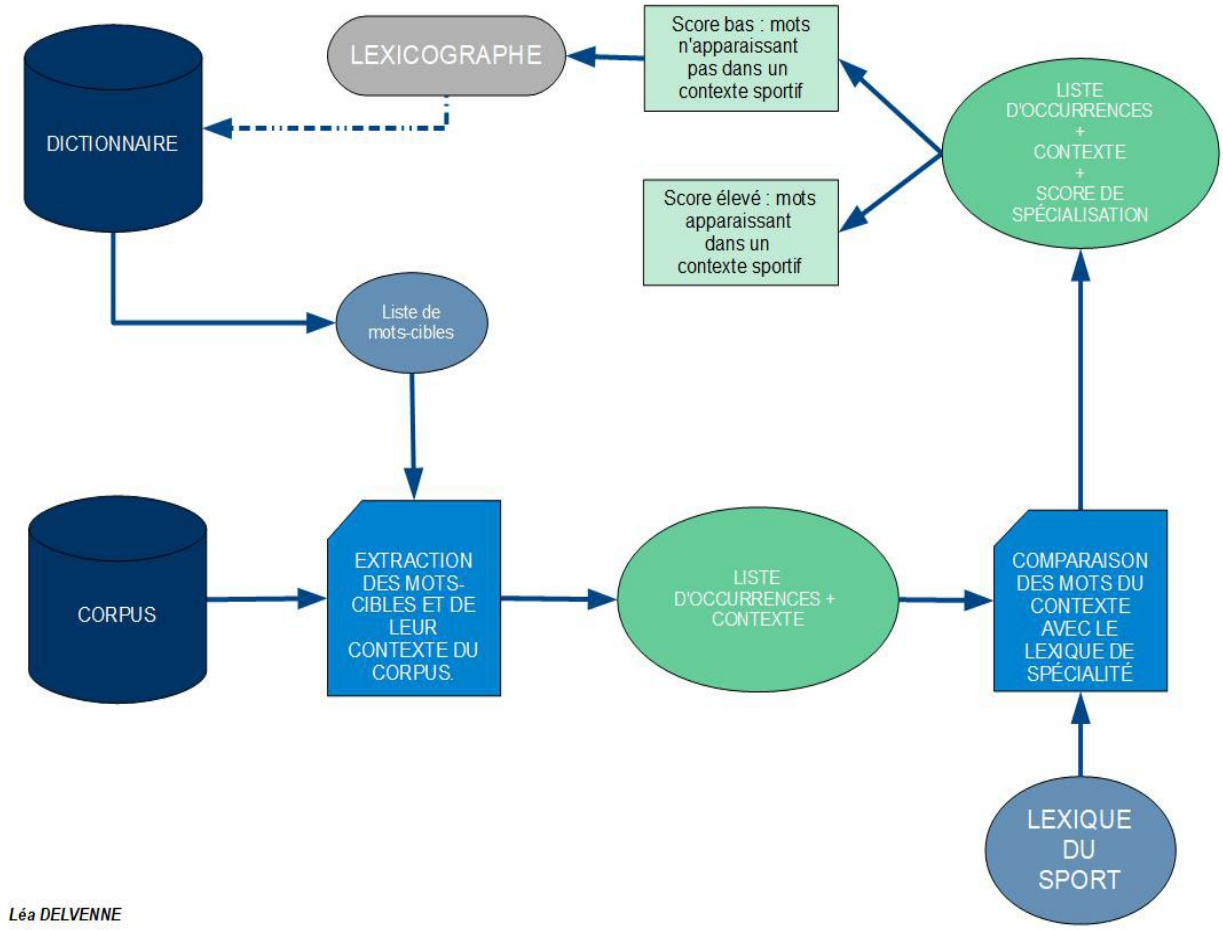



*Figure 1: schéma détaillant le processus de détection de cas de déterminologisation appliqué à la lexicographie.*

# CHAPITRE II : PRÉSENTATION DES RESSOURCES

## I. GLAWI

Dans cette partie, nous introduisons le dictionnaire GLAWI, que nous utilisons pour extraire des mots-cibles relevant du domaine du sport que nous souhaitons ensuite observer en contexte. Nous nous servons pour cela des marques de domaine, qui permettent de savoir de quel domaine relève un emploi d'un mot (cf. partie I. 2. de ce chapitre, p.22).

### 1. Présentation générale

GLAWI est un grand dictionnaire du Français disponible en téléchargement libre sur le site de CLLE-ERSS. Il a été créé par Franck Sajous en collaboration avec Basilio Calderone et Nabil Hatout (Sajous et Hatout, 2015). Le contenu de cette ressource est issu du Wiktionnaire, la version française du Wiktionary. GLAWI est encodé au format XML. Il est composé de 1 341 410 articles dans lesquels figurent les éléments suivants :

- mots simples, mots composés et locutions ;
- formes fléchies et leur lemme ;
- étymologie des mots ;
- prononciations, au format API ;
- définitions (gloses et exemples) ;
- traductions ;
- relations sémantiques ;
- relations morphologiques ;
- variantes orthographiques.[6]

La structure XML d'un article de GLAWI organise les informations de manière hiérarchisée, ce qui permet d'accéder facilement à celles dont nous avons besoin pour ce travail. Nous cherchons des mots employés dans le domaine sportif et voulons distinguer les mots monosémiques, apparaissant uniquement dans des emplois sportifs, des mots polysémiques, qui ont un sens relevant du domaine sportif et dont une ou plusieurs définitions relèvent aussi du langage général. Nous avons pour cela besoin d'accéder à leur(s) définition(s), mais également aux informations relatives aux usages de cette définition. Dans GLAWI, les sections *<definition>* se situent dans la balise *<POS>*. Chaque section *<definition>* comprend une sous section *<gloss>* dans laquelle se trouvent les informations que nous recherchons, à savoir les marques lexicographiques

6 http://redac.univ-tlse2.fr/lexiques/glawi.html

et le texte de la définition. Ainsi, chaque définition possède sa balise *<definition>* et peut se voir attribuer différents types de marques lexicographiques. Nous accédons donc aux informations recherchées en extrayant automatiquement les mots dont la ou l'une des sections *<definition>* contient une marque de domaine relevant du sport. La figure 2 (voir page suivante) permet de visualiser la structure d'un article de GLAWI. Nous y voyons les principales balises enfants directs de l'élément *<article>* ainsi que les enfants de l'élément *<POS>*. Nous expliquons dans la partie suivante comment repérer des marques de domaine et comment nous procédons pour extraire les mots qui nous intéressent.

## 2. Les marques de domaine dans GLAWI

GLAWI compte 388 noms de domaines différents, parmi lesquels nous pouvons relever par exemple BIOLOGIE, GÉOGRAPHIE, MUSIQUE, RELIGION, etc. Les marques de domaine ayant trait au sport retiennent notre attention.

La figure 2 est un extrait de l'article correspondant au mot *moniteur.* Ce dernier comprend douze définitions. La ligne en rouge correspond à la manière dont une marque de domaine est déclarée dans GLAWI. La ligne en vert correspond à un autre exemple de marque lexicographique. Nous avons également inséré une troisième définition pour montrer comment se présente une définition sans marque lexicographique.

```
<article>
  <title>moniteur</title>
  <pageId>32446</pageId>
  <text>
    <pos type="nom" lemma="1" locution="0" gender="m" number="s">
      <definitions>
        <definition>
          <gloss>
            <labels>
                <label type="diachronic" value="vieilli"/>
            </labels>
            <txt>Pédagogue, mentor (anglais moderne coach).</txt>
          </gloss>
          <example>
            <txt>Mentor fut le moniteur de Télémaque.</txt>
          </example>
        </definition>
        <definition>
          <gloss>
            <txt>Étudiant chargé d'enseigner sous forme de travaux dirigés, de cours pratiques sous l'autorité d'un
enseignant.</txt>
          </gloss>
          <example>
              <txt>Moniteur de l'enseignement supérieur.</txt>
          </example>
        </definition>
        <definition>
          <gloss>
            <labels>
                <label type="domain" value="sport"/>
            </labels>
            <txt>Personne chargée d'enseigner la pratique de certains sports.</txt>
          </gloss>
          <example>
              <txt>Moniteur de plongée sous-marine.</txt>
          </example>
        </definition>
      </definitions>
    </pos>
  </text>
</article>
```

*Figure 2: extrait de l'article du mot moniteur dans GLAWI.*

Les marques lexicographiques témoignent d'une certaine vision du monde. De Bessé (2000 : 187) met en évidence la dimension arbitraire de ces marques et la difficulté pour l'être humain de mettre au point une classification rigoureuse des emplois de la langue. Il rappelle en effet que les domaines morcellent les connaissances en de nombreuses rubriques qui

s'interpénètrent et sont souvent difficiles à délimiter. Sélectionner les marques se rapportant au sport implique de mener une réflexion sur ce qui relève, ou non du sport. Certaines personnes considèrent par exemple que la chasse ou la pêche sont des sports, d'autres non. Il est important de garder à l'esprit que le travail sur la sélection des marques de domaines sportifs est le fruit d'un choix qui peut varier d'une personne à l'autre. Nous sommes parti d'une définition trouvée dans le Larousse en ligne[7] :

> *« Ensemble des exercices physiques se présentant sous forme de jeux individuels ou collectifs, donnant généralement lieu à compétition, pratiqués en observant certaines règles précises ».*

*Figure 3: Définition du mot "sport" dans le Larousse en ligne.*

Nous avons décidé d'interpréter cette définition de manière assez large et d'inclure la chasse et la pêche dans la liste des domaines sportifs. Nous extrayons automatiquement une première liste composée de tous les noms de domaines existant dans GLAWI. Après une sélection manuelle, nous obtenons une seconde liste composée uniquement de domaine sportifs (tableau 1).

| | | | |
|---|---|---|---|
| aïkido | cricket | judo | ski de fond |
| alpinisme | cyclisme | motocyclisme | snowboard |
| arts martiaux | danse | musculation | sport |
| athlétisme | danses | natation | sports |
| automobile | équitation | navigation | sports de combat |
| badminton | escalade | patinage | sports de glisse |
| baseball | escrime | pêche | surf |
| basket-ball | football | pelote | tauromachie |
| billard | golf | pétanque | tennis |
| bowling | gymnastique | planche à neige | tennis de table |
| boxe | handball | planche à roulettes | volley-ball |
| chasse | hockey | rugby | |
| course à pied | jonglerie | ski alpin | |

*Tableau 1 : Liste des domaines sportifs extraits de GLAWI*

Ces domaines marquent 3132 mots en tout, que nous avons extraits automatiquement, ainsi que leurs définitions. Le tableau 2 indique la répartition de ces mots selon les domaines.

| MARQUE | NB ENTRÉES | MARQUE | NB ENTRÉES | MARQUE | NB ENTRÉES | MARQUE | NB ENTRÉES |
|---|---|---|---|---|---|---|---|
| aïkido | 6 | cricket | 5 | Judo | 23 | ski de fond | 11 |
| alpinisme | 42 | cyclisme | 80 | motocyclisme | 42 | snowboard | 20 |
| arts martiaux | 47 | danse | 110 | musculation | 28 | sport | 552 |
| athlétisme | 33 | danses | 12 | natation | 36 | sports | 35 |
| automobile | 159 | équitation | 314 | navigation | 50 | sports de combat | 9 |

7 http://www.larousse.fr/dictionnaires/francais/sport/74327?q=sport#73493

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| badminton | 6 | escalade | 20 | patinage | 11 | sports de glisse | 90 |
| baseball | 65 | escrime | 87 | pêche | 597 | surf | 11 |
| basket-ball | 7 | football | 78 | pelote | 26 | tauromachie | 37 |
| billard | 47 | golf | 44 | pétanque | 18 | tennis | 66 |
| bowling | 5 | gymnastique | 69 | planche à neige | 3 | tennis de table | 10 |
| boxe | 54 | handball | 4 | planche à roulettes | 7 | volley-ball | 1 |
| chasse | 504 | hockey | 5 | rugby | 61 | | |
| course à pied | 18 | jonglerie | 15 | ski alpin | 25 | | |

*Tableau 2: les marques de domaine relatives au sport et le nombre de mots marqués celles-ci.*

Notons qu'en calculant le nombre total des entrées pour chaque nom de domaine, nous obtenons un résultat supérieur au nombre de mots marqués. Cela est dû au fait que certains mots possèdent plusieurs noms de domaine. Par exemple, le mot *smash* est référencé comme employé dans les domaines du TENNIS, du TENNIS DE TABLE, du BADMINTON, du VOLLEY-BALL ET du BASKET-BALL.

### 3. Utilité pour le projet

Grâce à GLAWI nous construisons une liste de mots dont la ou l'une des définitions relève d'un domaine du sport. Nous pouvons à présent étudier ces mots en contexte pour voir si ce dernier a un impact sur l'emploi du mot. Nous détaillons dans le chapitre III, partie II. 1. (cf. p. 33) comment nous sélectionnons certains mots de cette liste pour tester notre hypothèse.

## II. LE CORPUS LE MONDE

### 1. Présentation générale

Nous disposons par ailleurs d'un corpus d'articles extraits du journal *Le Monde*. Distribué par ELRA, ce corpus contient des articles publiés entre 1991 et 2000. Dans le cadre de notre projet de recherche, nous nous servons d'un échantillon contenant les articles des années 1999-2000.

Le corpus est disponible dans deux formats différents. Le fichier initial est en texte brut (.txt) (cf. fig. 4 p.26), et une phase d'analyse syntaxique effectué par l'analyseur Talismane (Urieli, 2013) fournit un second fichier au format .tal (cf. fig. 5 p.26). L'avantage de disposer d'une version analysée réside dans la possibilité de travailler avec des lemmes, plutôt que de devoir récupérer systématiquement toutes les formes fléchies d'un mot. Par exemple, si nous souhaitons analyser les occurrences du mot *attaquant*, avec le texte brut il nous faudrait indiquer dans la liste de termes à analyser *attaquant/attaquante/attaquants/attaquantes*, pour que toutes les formes fléchies soient détectées. Avec le format .tal, toutes les formes fléchies du mot sont regroupées sous le lemme *attaquant*. Ainsi, dans la liste de termes à analyser et dans le lexique spécialisé,

nous n'avons qu'à indiquer le lemme pour que toutes les formes soient détectées. D'autre part, cela nous fournit des indications sur la catégorie grammaticale des mots du corpus.

<article id="LM10-d682668p4" rub="SPA">

INAUGURÉE dans une certaine gaieté avec le lancement de l'euro, l'année 1999 se termine dans une méditation songeuse, à Seattle, sur les avantages et les inconvénients de la mondialisation. D'une monnaie encore largement virtuelle, l'euro, au bon vieux roquefort, l'année a donc balancé entre européisme convaincu et antiaméricanisme rampant. Entre-temps, le drame du Kosovo s'est noué. 1999 restera marquée par les images de réfugiés kosovars cherchant à fuir leur pays noyé sous la neige, pourchassés par leurs tourmenteurs serbes. […]

*Figure 4: exemple d'un article au format .txt*

<article id="LM10-d682668p4" rub="SPA">

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | INAUGURÉE_ | | VPP | _ | _ | 17 | mod | 17 | mod | 100,00 | 5,47 | 80,11 |
| 2 | dans | dans | P | P | _ | 1 | mod | 1 | mod | 100,00 | 95,65 | 94,39 |
| 3 | une | une | DET | DET | g=f\|n=s | 5 | det | 5 | det | 100,00 | 98,61 | 99,82 |
| 4 | certaine | certain | ADJ | adj | g=f\|n=s | 5 | mod | 5 | mod | 100,00 | 96,39 | 99,42 |
| 5 | gaieté | gaieté | NC | nc | g=f\|n=s | 2 | prep | 2 | prep | 100,00 | 91,34 | 99,79 |
| 6 | avec | avec | P | P | _ | 1 | mod | 1 | mod | 100,00 | 97,20 | 93,70 |
| 7 | le | le | DET | DET | g=m\|n=s | 8 | det | 8 | det | 100,00 | 97,76 | 99,90 |
| 8 | lancement | lancement | NC | nc | g=m\|n=s | 6 | prep | 6 | prep | 100,00 | 98,29 | 99,85 |
| 9 | de | de | P | P | _ | 8 | dep | 8 | dep | 93,35 | 97,55 | 99,75 |
| 10 | l' | l' | DET | DET | n=s | 11 | det | 11 | det | 100,00 | 97,11 | 99,86 |
| 11 | euro | euro | NC | nc | g=m\|n=s | 9 | prep | 9 | prep | 100,00 | 79,92 | 99,95 |
| 12 | , | , | PONCT | PONCT | _ | 17 | ponct | 17 | ponct | 100,00 | 91,75 | 99,74 |
| 13 | l' | l' | DET | DET | n=s | 14 | det | 14 | det | 100,00 | 97,34 | 97,04 |
| 14 | année | année | NC | nc | g=f\|n=s | 17 | suj | 17 | suj | 100,00 | 99,49 | 99,56 |
| 15 | 1999 | _ | NC | _ | _ | 14 | mod | 14 | mod | 100,00 | 75,13 | 97,34 |
| 16 | se | se | CLR | CLR | n=p\|p=3 | 17 | aff | 17 | aff | 100,00 | 80,83 | 98,65 |
| 17 | termine | terminer | V | v | n=s\|p=13\|t=pst0 | | root | 0 | root | 100,00 | 95,13 | 99,45 |

*Figure 5: exemple d'article au format .tal.*

## 2. Description du contenu

Avant chaque article, une balise contient des informations d'identification et le nom de la

rubrique à laquelle cet article appartient. Cette balise se présente sous la forme suivante :

<article id="LM10-d682818p2" rub="SPO">

*Figure 6: exemple de balise d'un article du corpus Le Monde.*

Il y a 24 rubriques différentes, listées dans le tableau 3. Le nombre figurant sur leur droite correspond au nombre d'articles référencés pour la rubrique. La rubrique sport est notée SPO.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AGE | 479 | ECO | 18 | LIV | 617 | SOC | 1143 |
| ART | 1531 | ENT | 2303 | MDE | 529 | SPA | 703 |
| AUJ | 150 | ETR | 10 | MIA | 313 | SPO | 621 |
| CAR | 258 | FRA | 1114 | POC | 125 | TEL | 1062 |
| COM | 931 | HOR | 863 | QUO | 27 | TER | 329 |
| DER | 446 | INT | 2098 | SCI | 289 | UNE | 515 |

*Tableau 3: liste des rubriques des articles du corpus Le Monde.*

Le nom des rubriques nous permet de séparer automatiquement les articles en deux fichiers distincts : dans un fichier, un premier sous-corpus constitué des articles de sport (désormais LMsport), dans un autre fichier un second sous-corpus contenant le reste des articles (désormais LMautre). L'intérêt d'une telle séparation est que nous pouvons ainsi manipuler les articles de sport indépendamment des autres articles. Le corpus entier sera désormais désigné par LMglob.

Le corpus LMglob se compose de 16474 articles. 621 appartiennent à la rubrique sport (LMsport représente donc 3,77% de LMglob). Les 15853 restants forment donc LMautre qui représente 96,23% de LMglob.

## 3. Utilité pour le projet

Nous avons besoin d'observer des mots dans leur contexte pour voir si celui-ci suffit à repérer automatiquement un emploi spécialisé. Le corpus *Le Monde* nous fournit des articles de la rubrique sport dans lesquels se trouve une part de vocabulaire propre au sport, mais nous avons également des articles traitant d'autres sujets dans lesquels le vocabulaire est différent. Grâce aux rubriques, nous pouvons établir deux sous corpus qui nous permettent de définir un vocabulaire spécifique au sport pour établir un lexique spécialisé. Les détails de constitution de ce lexique sont décrit dans le chapitre III (cf. partie II. 2. p. 36). Nous nous servons également des sous-corpus

pour comparer les résultats obtenus pour les articles de sport et ceux obtenus pour le reste des articles. Sachant que les mots analysés ont plus de chance d'apparaître dans un contexte sportif quand ils sont dans un article de sport, s'il s'avère que les emplois sportifs apparaissent plutôt dans les articles de sport et les emplois non sportifs plutôt dans le reste des articles nous pourrons en déduire que le contexte permet déterminer automatiquement l'emploi d'un mot.

# CHAPITRE III : MÉTHODOLOGIE

## I. FONCTIONNEMENT DU PROGRAMME

### 1. Principe

Nous expliquons maintenant la procédure que nous suivons pour vérifier la validité de l'hypothèse émise. Partant de là, nous imaginons un programme qui recherche en corpus des occurrences de mots-cibles, puis qui analyse le contexte dans lequel ces occurrences apparaissent pour calculer la proportion de mots du sport qui s'y trouvent. Les mots-cibles, dont nous décrivons les modalités de sélection dans la partie II. 1. de ce chapitre (cf. p. 33), sont issus de GLAWI. Les mots du sport sont contenus dans un lexique qui doit refléter le vocabulaire du domaine de la façon la plus précise possible. Nous expliquons dans la partie II. 2. de ce chapitre comment nous construisons ce lexique à la fois à partir de GLAWI et à partir du sous corpus LMsport (cf. p. 36).

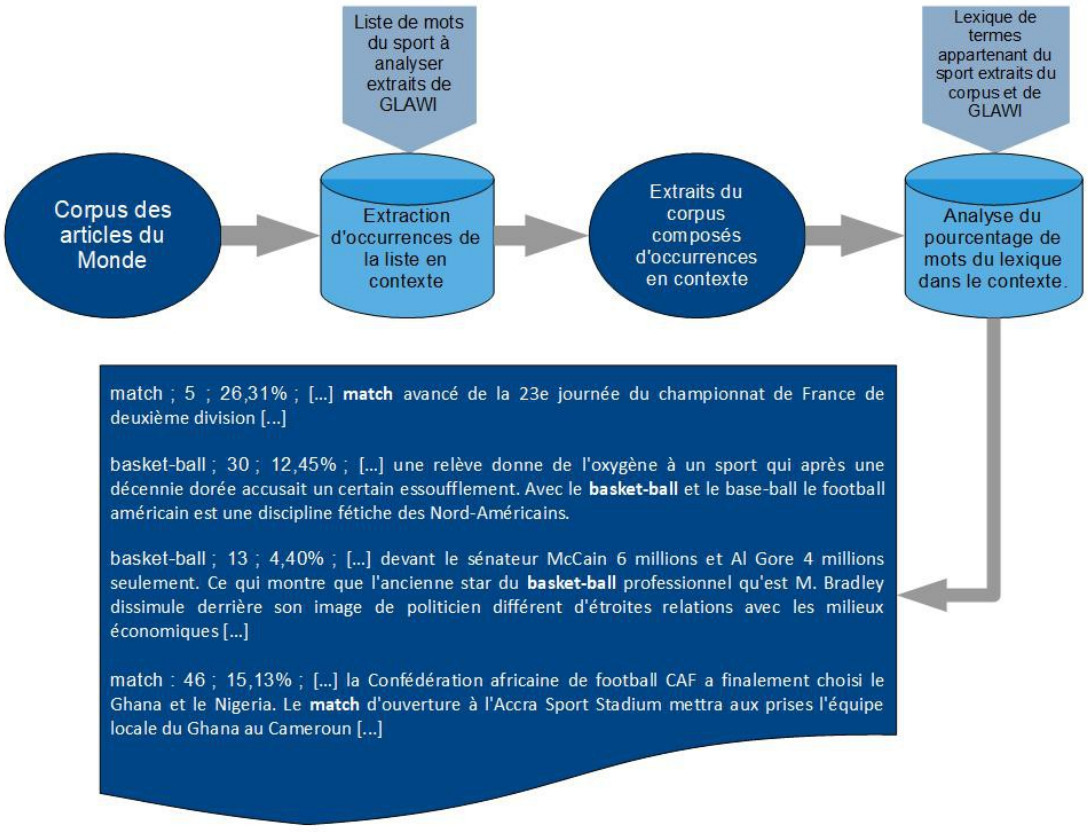



*Figure 7 : schéma explicitant le fonctionnement du programme de détection des termes despécialisés en fonction de la nature de leur contexte.*

## 2. Entrée et sortie

La figure 7 (p. 29) permet de visualiser le principe de fonctionnement du programme à implémenter. Les corpus en entrée sont les corpus LMglob, LMsport ou LMautre étiquetés par Talismane. Le fichier de sortie attendu est un fichier .csv dans lequel apparaissent les occurrences, le nombre de mots du lexique du sport trouvés dans le contexte, le pourcentage que cela représente par rapport au reste de la fenêtre contextuelle et enfin un extrait du contexte pour avoir une idée de la façon dont l'occurrence est employée. Le tableau 4 (p. 31) vient en complément du schéma pour illustrer le type de fichier de sortie attendu. Les paramètres du tableau, à savoir le nombre et le pourcentage de mots du lexique sont expliqués dans la partie I. 3. de ce chapitre (cf. p. 32). Notons que, par souci de place, l'ensemble de la fenêtre contextuelle (qui correspond à l'article en entier) n'apparaît pas dans le tableau, et ce pour tous les tableaux présentant des exemples de fichiers de sortie. Par conséquent, les nombres des colonnes 2 et 3 ne correspondent pas à l'extrait de contexte visible. Dans le contexte, le mot en rouge est le mot-cible et les mots en gras sont les mots appartenant au lexique. Le texte 6 (p. 31) est un exemple de traitement pour un article entier. Il correspond au contexte dans lequel apparaît la cinquième occurrence du tableau (demi-finale).

| TERME PROJETÉ | NOMBRE DE MOTS DU LEXIQUE | POURCENTAGE DE MOTS DU LEXIQUE | OCCURRENCE EN CONTEXTE. |
|---|---|---|---|
| footballeur | 59 | 12,40% | Certains exercent leur métier au sein des meilleures **ligues** professionnelles européennes. Si cette prédominance témoigne de l'intérêt grandissant porté pour les **footballeurs**, africains la situation n'est pas sans poser problème à chaque fois que se déroule la CAN |
| match | 29 | 23,10% | [...] l'Allemagne a obtenu trois premières places au point de conférer à ces **championnats** du monde une image de **match** franco-allemand. La France a pu compter sur ses féminines Félicia Ballanger, bien sûr, dont c'était le dernier **mondial vitesse**[...] |
| hippodrome | 12 | 4,36% | L'ANNONCE surprise par le maire de Paris, Jean Tiberi (RPR), vendredi 7 janvier, de son intention de transformer l'un des trois **hippodromes** qui jouxtent la capitale en un immense espace de détente et de promenade à la disposition des Parisiens (Le Monde daté 9-10 janvier) a déclenché un tollé dans les milieux des **courses**. |
| match | 5 | 26,31% | [...] **match** avancé de la 23e journée du **championnat** de France de **deuxième division** [...] |
| demi-finale | 11 | 26,80% | [...] l'**équipe** de France de **rugby** a réalisé le plus grand exploit de son histoire : **battre** 43 -31 en **demi-finales** de la **Coupe** du monde son homologue néo-zélandaise, grande favorite de la **compétition**. |
| basket-ball | 5 | 50,00% | [...] une relève donne de l'oxygène à un **sport** qui après une décennie dorée accusait un certain essoufflement. Avec le **basket-ball** et le **base-ball** le **football** américain est une **discipline** fétiche des Nord-Américains. |
| basket-ball | 13 | 4,40% | [...] devant le sénateur McCain 6 millions et Al Gore 4 millions seulement. Ce qui montre que l'ancienne star du **basket-ball** professionnel qu'est M. Bradley dissimule derrière son image de politicien différent d'étroites relations avec les milieux économiques [...] |
| match | 46 | 15,13% | [...] la Confédération africaine de **football** CAF a finalement choisi le Ghana et le Nigeria. Le **match** d'ouverture à l'Accra **Sport** Stadium mettra aux prises l'**équipe** locale du Ghana au Cameroun [...] |
| boxeur | 5 | 2,00% | Une vision assez drôle de certains aspects du rituel - le jeune garçon bar-mitsva apparaît lors de sa fête en peignoir de **boxeur** sous l'air de la musique de Rocky III [...] |

*Tableau 4: exemple de fichier attendu en sortie du programme d'extraction.*

<article id="LM10-d682736p1" rub="SPA">

Dimanche 31 octobre, à Twickenham, l'**équipe** de France de **rugby** a réalisé le plus grand exploit de son histoire : **battre** (43-31), en **demi-finales** de la **Coupe** du monde, son homologue néo-zélandaise, grande favorite de la **compétition**. Les Bleus ne parviendront pas à rééditer une **performance** de même niveau une semaine plus tard, en **finale**, laissant les Australiens emporter leur **deuxième** trophée Webb-Ellis.

*Texte 6: exemple d'un article du Monde avec le terme-cible (en rouge) et les mots du lexique spécialisé (en gras).*

### 3. Calcul du score de spécialisation

Le score de spécialisation permet d'établir automatiquement si l'emploi est spécialisé ou non. L'objectif est de fixer une valeur seuil au dessus de laquelle l'occurrence est considérée comme relevant d'un emploi sportif. Nous avons choisi d'effectuer un calcul de pourcentage : le programme compte dans la fenêtre contextuelle le nombre de mots qui font partie du lexique spécialisé (hors mot-cible) et le nombre total de mots. Pour calculer ce pourcentage, nous avons décidé de ne prendre en compte que les mots lexicaux, c'est-à-dire ceux qui sont susceptibles de participer à créer un contexte sportif. En effet, en observant la classe grammaticale des mots du lexique, nous nous sommes aperçu qu'il s'agissait exclusivement de noms, de verbes ou d'adjectifs. Cela signifie que si nous comptons les mots grammaticaux dans le nombre total de mots de la fenêtre, nous prenons en compte, dans le calcul, des mots qui, de par leurs propriétés grammaticales, ne participeront jamais à créer un contexte sportif, tout du moins avec le lexique établi. Nous limitons ainsi la possibilité de faire tendre le résultat du score vers 100%. Pour équilibrer le score, lors de l'analyse du corpus, nous nous servons de la quatrième colonne du fichier étiqueté (qui correspond à la nature des lemmes) pour éliminer les mots grammaticaux. Cela signifie que ces derniers ne sont pas comptabilisés dans le nombre total de mots de la fenêtre contextuelle. Notre calcul du pourcentage de mots du lexique dans une fenêtre contextuelle est le suivant :

$$score\ de\ spécialisation = \frac{nombre\ de\ mots\ appartenant\ au\ lexique}{nombre\ de\ mots\ lexicaux} \times 100$$

La taille de la fenêtre contextuelle correspond à la taille de l'article dans lequel l'occurrence d'un mot a été trouvée. Les articles de presse sont généralement plutôt courts et centrés sur un sujet, ce qui limite le risque d'incohérences. En effet, si un article traite d'un événement sportif, il y a peu de chance que le sujet change au cours de l'article, donc a priori le type de contexte restera sensiblement le même tout du long. De même si l'article est centré sur un autre sujet. Parmi les ajustements à faire pour avoir la meilleure configuration nous avons envisagé de faire des tests sur plusieurs tailles de fenêtres contextuelle (une phrase, trois phrases, un paragraphe,...) pour voir dans quels cas nous obtenions les meilleurs résultats.

## II. CONSTRUCTION DES OBJETS DU PROGRAMME

Nous avons besoin de constituer deux listes de termes issus du domaine du sport (cf. II. 1. de ce chapitre, p. 33). Il s'agit de définir des mots-cibles à rechercher en contexte. Dans un premier temps nous devons nous assurer que les emplois sportifs apparaissent bien dans des contextes sportifs et que les contextes sportifs sont bien détectés par le programme. Nous choisissons pour cela des mots monosémiques, dont les emplois sont censés être uniquement liés au sport. Cette liste de mots-cibles permettra d'observer les résultats obtenus pour des mots supposés sans ambiguïté sémantique, et d'identifier le type de difficultés qui peuvent se manifester même

lorsqu'il n'y a pas d'ambiguïté sémantique. Si notre hypothèse se vérifie, les scores de pourcentage doivent révéler un emploi sportif pour chaque occurrence. Nous les extrayons en cherchant dans GLAWI les mots marqués SPORT ou autre nom de domaine lié au sport et dont l'article ne contient qu'une balise *<definition>.* Dans un second temps, nous devons observer le traitement que fait le programme de mots dont la polysémie est avérée. Un mot polysémique est un mot qui a différents sens, spécialisés ou non. Nous les extrayons en cherchant dans GLAWI les mots dont l'article a plusieurs balises *<definition>* et dont l'une de ces définitions a une marque de domaine liée au sport. Nous nous servons de cette seconde liste pour vérifier que les emplois non sportifs apparaissent bien dans des contextes pauvres en mots spécialisés, et que le pourcentage moyen pour les emplois sportifs se distingue bien du pourcentage moyen pour les emplois non sportifs. Nous voulons avoir une idée du type de résultats que nous pourrions obtenir face à un vrai cas de despécialisation. Pour cette seconde phase, nous avons besoin d'une liste de mots ayant un emploi dans le sport.

Nous avons également besoin d'identifier automatiquement un contexte sportif. Pour cela, nous choisissons de construire un lexique de mots faisant partie du vocabulaire sportif. Il est construit à partir du corpus LMsport et des mots de GLAWI (cf. II. 2. de ce chapitre, p. 36). Ces mots seront comparés à ceux formant le contexte du terme étudié.

## 1. Constitution des listes de termes monosémiques et polysémiques du sport (mots-cibles).

Nous constituons deux listes. La première se compose de mots monosémiques, nous l'appelons monoSport. La seconde se compose de mots polysémiques, nous l'appelons polySport.

### a. Constitution de la liste monoSport

| CLASSE GRAMMATICALE | NOMBRE DE MOTS | POURCENTAGE |
|---|---|---|
| Noms | 1270 | 89,63% |
| Verbes | 100 | 7,06% |
| Adjectifs | 32 | 2,26% |
| Interjections | 7 | 0,49% |
| Adverbes | 6 | 0,42% |
| Noms propres | 1 | 0,07% |
| Suffixes | 1 | 0,07% |

*Tableau 5: répartition des différentes classes grammaticales des mots monosémiques de GLAWI*

Nous constituons une pré-liste qui contient tous les mots monosémiques de GLAWI dont la définition porte une marque de domaine en rapport avec le sport. Elle représente 1417 entrées du dictionnaire. Parmi ces entrées, la grande majorité sont des noms. Le détail de la répartition des mots selon la classe grammaticale se trouve dans le tableau 5.

Parmi les mots que nous venons d'isoler, nous voulons des termes qui apparaissent dans LMglob, afin d'avoir des occurrences à analyser. Parmi les 1417 entrées initiales, il y en a 46 qui sont représentées dans le corpus (voir tableau 6).

| ENTRÉE | POS | FREQ | ENTRÉE | POS | FREQ | ENTRÉE | POS | FREQ |
|---|---|---|---|---|---|---|---|---|
| match | NOM | 145 | péniche | NOM | 3 | aïkido | NOM | 1 |
| demi-finale | NOM | 59 | alpinisme | NOM | 2 | beach-volley | NOM | 1 |
| footballeur | NOM | 19 | amphétamine | NOM | 2 | braquet | NOM | 1 |
| basket-ball | NOM | 16 | bad | NOM | 2 | capoeira | NOM | 1 |
| bal | NOM | 12 | chaluter | VERBE | 2 | club-house | NOM | 1 |
| boxeur | NOM | 11 | demi-finaliste | NOM | 2 | coupé-cabriolet | NOM | 1 |
| hippodrome | NOM | 9 | handball | NOM | 2 | cricket | NOM | 1 |
| foot | NOM | 6 | jonglage | NOM | 2 | hippique | ADJ | 1 |
| buteur | NOM | 5 | judo | NOM | 2 | pluvier | NOM | 1 |
| cross-country | NOM | 4 | judoka | NOM | 2 | sampan | NOM | 1 |
| mercato | NOM | 4 | rugbyman | NOM | 2 | squash | NOM | 1 |
| super-G | NOM | 4 | tennisman | NOM | 2 | torero | NOM | 1 |
| autoradio | NOM | 3 | virevolter | VERBE | 2 | trappeur | NOM | 1 |
| cyclomoteur | NOM | 3 | volley-ball | NOM | 2 | tétras | NOM | 1 |
| endurance | NOM | 3 | autoallumage | NOM | 1 | | | |
| penalty | NOM | 3 | avant-centre | NOM | 1 | | | |

*Tableau 6 : les entrées de GLAWI ayant au moins une occurrence dans le corpus Le Monde classées par ordre de fréquence.*

Parmi ces entrées, nous effectuons une seconde sélection, en faisant attention que les mots soient représentés à la fois dans LMglob et dans LMsport. Nous souhaitons en effet savoir si les résultats diffèrent d'un sous-corpus à l'autre (si les pourcentages sont plus hauts dans le corpus sport par exemple) et si ces différences sont liées au type de contexte dans lequel apparaissent les occurrences.

D'autre part, certains mots, qui peuvent apparaître dans des contextes ambigus, ne sont pas sélectionnés, car l'intérêt de cette liste est de nous permettre de savoir si les emplois sportifs ont un *« profil type »* que l'on puisse exploiter pour les repérer. Nous constituons une liste de termes monosémiques pour avoir le moins d'ambiguïté possible sur le type de contexte récupéré, mais certains mots sont considérés comme monosémiques dans GLAWI et pourraient faire l'objet de modifications. Par exemple, le mot *bal* fait partie des mots monosémiques du sport car il est étiqueté *danse* et il apparaît douze fois dans le corpus. Or, une observation en corpus nous montre que les emplois les plus courants ont lieu dans des expressions du type *« ouvrir le bal »* ou

*« mener le bal »* comme dans cette phrase, tirée d'un article économique :

*« Depuis un an, l'Europe mène le bal des fusions et acquisitions. »*

Une fois ce processus de sélection effectué, nous obtenons une liste monoSport composée des termes-cible suivants :

| TERME | POS | FREQ LMglob | FREQ LMsport | TERME | POS | FREQ LMglob | FREQ LMsport |
|---|---|---|---|---|---|---|---|
| match | NOM | 145 | 74 | hippodrome | NOM | 9 | 0 |
| demi-finale | NOM | 59 | 50 | foot | NOM | 6 | 0 |
| footballeur | NOM | 19 | 5 | buteur | NOM | 5 | 4 |
| basket-ball | NOM | 16 | 11 | cross-country | NOM | 4 | 4 |
| boxeur | NOM | 11 | 3 | penalty | NOM | 0 | 3 |

*Tableau 7 : les dix termes issus de GLAWI sélectionnés pour constituer la liste monoSport.*

## b. Constitution de la liste polySport

Nous suivons une procédure similaire pour constituer la liste polySport. Les mots polysémiques de GLAWI représentent 1715 entrées du dictionnaire, dont les classes grammaticales se répartissent dans des proportions similaires à celles des mots monosémiques (cf. tableau 5).

Comme pour monoSport, nous cherchons des mots représentés dans le corpus, donc nous regardons la fréquence des mots de GLAWI dans LMglob, puis dans LMsport. Nous effectuons notre sélection en nous assurant de leur présence dans LMsport, afin d'augmenter la probabilité d'avoir des emplois sportifs. En effet, ces termes étant polysémiques, il est possible qu'ils soient fréquents dans le corpus global mais seulement dans un emploi général ou autre que sportif. Afin d'avoir des données similaires, nous souhaitons obtenir un fichier de sortie qui fasse une taille équivalente à celle du fichier obtenu avec les mots de monoSport, donc comprenant environ trois-cents occurrences. Nous prévoyons en effet une phase d'annotation manuelle et pour cette raison, nous ne voulons pas de données trop nombreuses. Nous excluons certains noms, comme *par* (qui, dans le golf, fait référence au *nombre conventionnel de coups qui sont nécessaires pour faire le parcours d'un trou* - selon GLAWI) ou *pas*, (qui désigne les pas de danse ou l'allure du cheval en équitation) d'une part parce qu'ils sont trop fréquents et d'autre part parce qu'ils peuvent aisément être mal étiquetés.

Atkins et Rundell (2008 : chapitre 8) insistent sur l'importance de bien distinguer l'homonymie de la polysémie lors de la rédaction d'un dictionnaire. *Par* et *pas* sont des exemples,

mais nous pensons également à *stade*, qui désigne à la fois l'infrastructure et une étape dans un processus. Il ne s'agit pas d'une déterminologisation, cependant, pour les besoins du programme, il est intéressant d'observer la manière dont les deux acceptions sont employées et s'il est possible de les distinguer automatiquement. D'ailleurs, il est important de préciser que nous n'avons pas spécialement cherché des mots déterminologisés pour tester l'hypothèse, mais, dans un premier temps, des mots ayant plusieurs emplois dans des contextes différents. Il s'agit dans un premier temps de s'assurer que le contexte permet de distinguer automatiquement emploi sportif et emploi non sportif. Nous décidons donc de ne pas éliminer *stade* pour l'expérience, puisque nous cherchons à savoir si les deux types d'emplois sont assez marqués pour être détectables, mais nous sommes conscient qu'une sélection plus rigoureuse mériterait d'être effectuée pour des recherches ultérieures. Une étape supplémentaire pourrait d'ailleurs consister à sélectionner un échantillon de dix termes despécialisés pour s'assurer de l'issue de l'hypothèse.

Nous choisissons ici encore 10 mots-cibles en prenant soin que le nombre total d'occurrences dans LMglob avoisine les 300.

| TERME | POS | FREQ LMglob | FREQ LMsport | TERME | POS | FREQ LMglob | FREQ LMsport |
|---|---|---|---|---|---|---|---|
| Stade | NOM | 70 | 18 | Saut | NOM | 24 | 7 |
| Arbitre | NOM | 42 | 13 | Peloton | NOM | 19 | 11 |
| Pénalité | NOM | 37 | 27 | Médaille | NOM | 17 | 6 |
| Challenger | NOM | 27 | 21 | Attaquant | NOM | 16 | 13 |
| Pompe | NOM | 24 | 1 | Marathon | NOM | 9 | 3 |

*Tableau 8 : les dix termes issus de GLAWI sélectionnés pour constituer la liste polysport.*

## 2. Constitution du lexique permettant d'identifier un contexte sportif

Le lexique spécialisé du sport permet de d'identifier les mots du contexte qui appartiennent au domaine du sport. Désormais nous l'appelleront LexSpo. Pour rassembler des mots spécifiques au sport, nous nous servons de deux sources différentes. D'abord, nous récupérons le vocabulaire spécifique au sous-corpus LMsport pour sélectionner les mots susceptibles de contribuer à créer un contexte sportif. Puis nous insérons tous les mots monosémiques étiquetés *sport* de GLAWI. La figure 8 illustre le processus suivi pour constituer le lexique LexSpo.

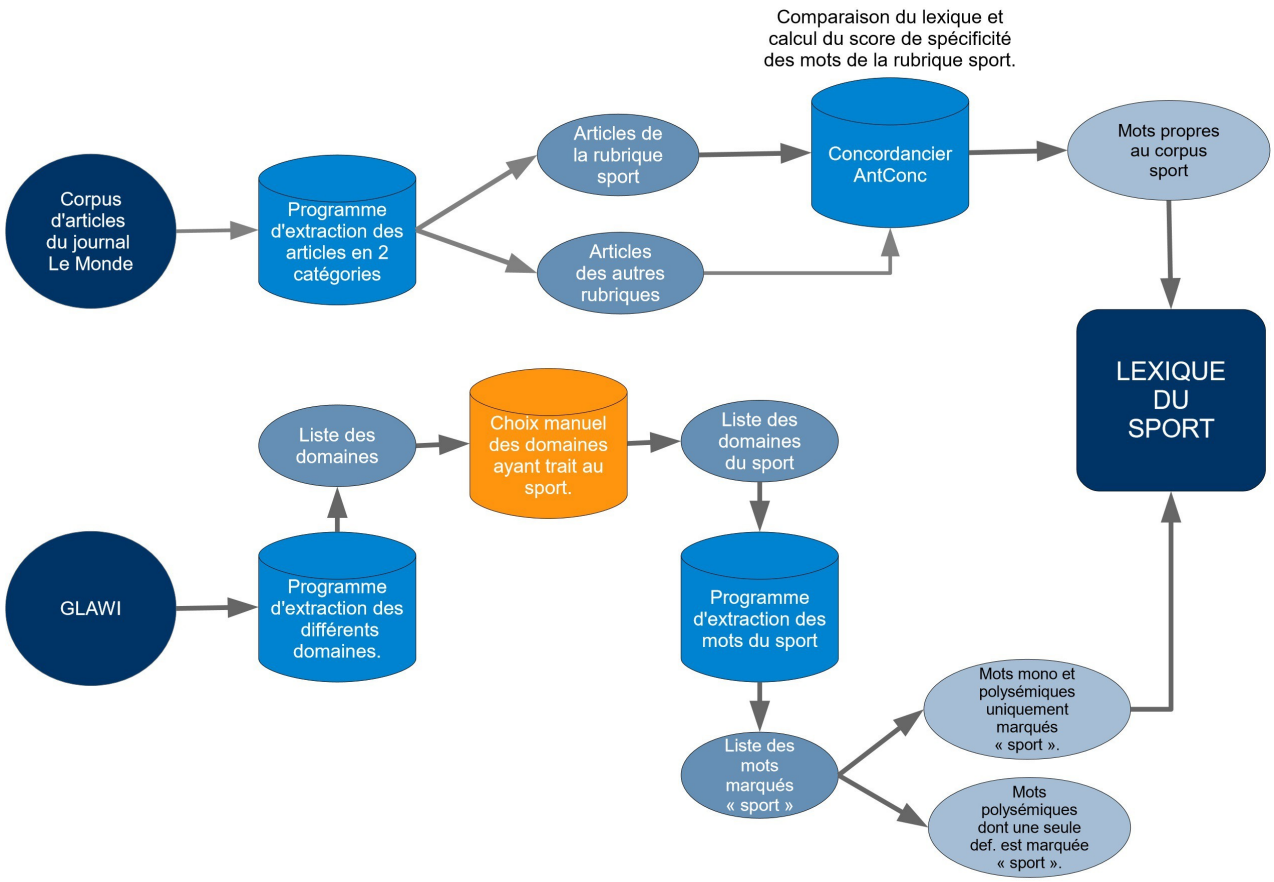



*Figure 8: schéma décrivant les étapes de constitution de LexSpo*

Nous postulons que le corpus LMSport, composé des articles de sport, possède un vocabulaire du sport que nous pouvons extraire pour constituer le lexique. Pour cela, nous utilisons le logiciel d'analyse de corpus AntConc (Anthony L., 2014)[8] et calculons le score de spécificité lexicale les lemmes de LMSport par rapport à LMAutre. AntConc propose deux calculs différents pour déterminer le score de spécificité ; le *khi-deux* et le *log-likelihood*. Une analyse plus approfondie de la façon dont ils fonctionnent, à la manière de celle menée par Kilgarriff (2001), permettrait de faire un choix éclairé en faveur de l'un ou l'autre. Nous avons choisi d'effectuer une comparaison manuelle des résultats de ces deux calculs et avons jugé qu'il n'y avait pas de grosses variations dans les résultats selon que nous utilisions un calcul ou l'autre. Nous choisissons le khi-deux, tout en étant bien conscient que le choix de la mesure pourrait être intégrée comme un paramètre à ajuster dans notre chaîne de traitement. Les mots sont alors classés dans l'ordre décroissant de leur score de spécificité, et nous choisissons ensuite manuellement les 250 premiers mots pertinents pour identifier un contexte sportif. Nous sélectionnons les mots lexicaux que nous identifions comme appartenant au domaine du sport. Il s'agit de vocabulaire lié à la compétition sportive (*match, vainqueur, tournoi, coupe, podium, classement*,...), aux sports d'équipe (*balle, ballon, championnat, maillot, but, buteur, mêlée,*...), à l'athlétisme (*athlétisme,*

8 AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Disponible à l'adresse URL suivante : http://www.laurenceanthony.net/

*coureur, marathon, piste*,...), à la voile (*bateau, régate, challenger, risée, démâtage,*...), etc. Les statistiques sur la classe grammaticale des mots de GLAWI liés au sport révèlent que la plupart des mots identifiés comme propres au domaine sportif, donc susceptibles de permettre de distinguer contexte sportif de contexte général, sont des noms, mais également des verbes et des adjectifs. Nous y ajoutons quelques noms propres ainsi que les sigles de noms d'équipes sportives ou de fédérations connues qui apparaissent dans la liste de spécificité, car lorsqu'une occurrence est un emploi sportif qui apparaît dans un contexte non sportif, il arrive régulièrement qu'il s'agisse d'un article traitant d'un sportif ou d'une équipe. Nous donnons quelques exemples de ce phénomène dans le chapitre IV (cf. I. p.41).

Nous obtenons un lexique LexSpo de 350 mots susceptibles de se retrouver dans un contexte sportif. Ces mots sont des noms à 70,5% ; il y a 15% d'adjectifs et 10% de verbes. Il y a également 4,5% de sigles comme OL, OM, PSG, FFR, FIFA, etc.

Notons que notre lexique contient peu de noms propres, à l'exception de certains noms d'équipes sportives comme le Paris Saint-Germain ou l'Olympique Lyonnais. Ce choix est justifié par le fait qu'une réflexion plus aboutie sur la nature des noms propres (noms de sportifs, noms d'équipes, noms de clubs,...) et leur pertinence dans le lexique serait nécessaire avant de décider d'en ajouter plus ou non. En effet, il faudrait mener une étude afin de savoir s'ils contribuent vraiment à obtenir des scores plus précis, ou si dans certains cas, ils n'augmentent pas le pourcentage dans un contexte ou cela n'a pas lieu d'être. Les noms propres que nous avons choisi d'ajouter sont des noms d'équipes sportives qui ont obtenu un haut score de spécificité. Nous avons décidé de ne pas mener cette réflexion pour le moment mais nous sommes conscients qu'il s'agit d'un aspect perfectible de notre méthodologie.

## III. ANNOTATION

À ce stade, nous pouvons calculer le nombre et le pourcentage de mots de LexSpo dans le contexte d'une occurrence d'un terme issu de monoSport ou polySport. L'objectif est d'obtenir un format de sortie dans lequel, pour une valeur seuil donnée, si le pourcentage est inférieur à la valeur seuil déterminéé, le résultat est 0 et si le pourcentage est supérieur à la valeur seuil le résultat est 1. 0 et 1 signifient ici « il ne s'agit pas / il s'agit d'un emploi sportif ». Le bon fonctionnement d'un tel programme dépend de l'ajustement de la valeur seuil. Pour faire cet ajustement, nous voulons effectuer plusieurs *runs* en modifiant le seuil à chaque *run*, et comparer les résultats obtenus à un fichier *gold*, annoté manuellement. Cette technique permet de savoir pour quels paramètres les résultats sont le plus proches d'une décision humaine.

Le tableau 9 illustre le type de fichiers que nous voulons obtenir en sortie. Il s'agit d'un

format .csv obtenu en exécutant le programme sans valeur seuil (voir tableau 4). Nous effectuons trois *runs* avec les mots de monoSport dans LMglob, LMsport et LMautre, puis nous répétons l'opération pour la liste polySport. Nous avons alors six fichiers de sortie. Puis, pour chacun des fichiers, nous ajoutons deux colonnes : s'agit-il d'un emploi sportif (c'est-à-dire, l'occurrence analysée renvoie-t-elle au sens sportif du mot) ? - 0 ou 1 – l'emploi apparaît-il dans un contexte sportif ? - 0 ou 1. Nous complétons ces colonnes pour chaque ligne du tableau.

Nous considérons comme emploi sportif toute acception dont le sens renvoie à une définition du mot qui se rapporte au sport. Un emploi non-sportif est alors une acception dont le sens ne renvoie pas à une définition se rapportant au sport. Par ailleurs, un contexte est considéré comme étant de type sportif lorsque le propos tenu se rapporte au sport. Il est de type non sportif lorsque le propos fait référence à autre chose que le sport. Par exemple :

> *« Elle illustre aussi le courage de Derek Redmond, qui termine sa course malgré une blessure à mi-parcours (Barcelone, 1992) et met en exergue le bonheur de la performance plus que celui de « la victoire à tout prix » avec l'équipe nigériane du* ***relais*** *quatre fois 100 mètres ivre de joie d'avoir obtenu une médaille de bronze (Barcelone). »* (tiré de l'article id="LM10-d686397p4" rub="COM")

> *« A Montpellier (Hérault), les enseignants et parents du collège des Aiguerelles devaient décider vendredi 21 , dans la soirée, des suites à donner à leur mouvement, entamé il y a quinze jours pour protester contre les agressions, les incivilités quotidiennes et le racket. Après dix journées de grève des enseignants, les parents avaient pris le* ***relais****, en début de semaine, en campant dans l'établissement. »* (tiré de l'article id="LM10-d686484p2" rub="SOC")

Dans l'exemple 1, le mot relais renvoie au sens dont la définition se rapporte au sport, il s'agit donc d'un emploi sportif. Il apparaît dans un contexte dont le propos se rapporte également au sport puisqu'il est question de performance sur une compétition. Il s'agit donc d'un contexte sportif. Dans l'exemple 2 en revanche, le mot relais renvoie au sens figuré, dont la définition ne se rapporte pas au sport. Il s'agit d'un emploi non sportif. Le contexte dans lequel il apparaît ne tient pas un propos relatif au sport, il est donc de type non sportif.

| TERME PROJETÉ | NOMBRE DE MOTS DU LEXIQUE | % DE MOTS DU LEXIQUE | OCCURRENCE EN CONTEXTE. | CONTEXTE SPORTIF (0/1) | EMPLOI SPORTIF (0/1) |
|---|---|---|---|---|---|
| footballeur | 59 | 12,40% | Certains exercent leur métier au sein des meilleures **ligues** professionnelles européennes. Si cette prédominance témoigne de l'intérêt grandissant porté pour les **footballeurs**, africains la situation n'est pas sans poser problème à chaque fois que se déroule la CAN | 1 | 1 |
| match | 29 | 23,10% | […] l'Allemagne a obtenu trois premières places au point de conférer à ces **championnats** du monde une image de **match** franco-allemand. La France a pu compter sur ses féminines Félicia Ballanger, bien sûr, dont c'était le dernier **mondial vitesse**[...] | 1 | 1 |
| hippodrome | 12 | 4,36% | L'ANNONCE surprise par le maire de Paris, Jean Tiberi (RPR), vendredi 7 janvier, de son intention de transformer l'un des trois **hippodromes** qui jouxtent la capitale en un immense espace de détente et de promenade à la disposition des Parisiens (Le Monde daté 9-10 janvier) a déclenché un tollé dans les milieux des **courses**. | 0 | 1 |
| match | 5 | 26,31% | […] **match** avancé de la 23e journée du **championnat** de France de **deuxième division** [...] | 1 | 1 |
| demi-finale | 11 | 26,80% | [...] l'**équipe** de France de **rugby** a réalisé le plus grand exploit de son histoire : **battre** 43 -31 en **demi-finales** de la **Coupe** du monde son homologue néo-zélandaise, grande favorite de la **compétition**. | 1 | 1 |
| basket-ball | 5 | 50,00% | […] une relève donne de l'oxygène à un **sport** qui après une décennie dorée accusait un certain essoufflement. Avec le **basket-ball** et le **base-ball** le **football** américain est une **discipline** fétiche des Nord-Américains. | 1 | 1 |
| basket-ball | 13 | 4,40% | […] devant le sénateur McCain 6 millions et Al Gore 4 millions seulement. Ce qui montre que l'ancienne star du **basket-ball** professionnel qu'est M. Bradley dissimule derrière son image de politicien différent d'étroites relations avec les milieux économiques [...] | 0 | 1 |
| match | 46 | 15,13% | […] la Confédération africaine de **football** CAF a finalement choisi le Ghana et le Nigeria. Le **match** d'ouverture à l'Accra **Sport** Stadium mettra aux prises l'**équipe** locale du Ghana au Cameroun [...] | 1 | 1 |
| boxeur | 5 | 2,00% | Une vision assez drôle de certains aspects du rituel - le jeune garçon bar-mitsva apparaît lors de sa fête en peignoir de **boxeur** sous l'air de la musique de Rocky III [...] | 0 | 1 |

*Tableau 9: exemple d'un fichier de sortie annoté.*

Nous pouvons comparer ces fichiers annotés aux *runs* fournis par le programme et définir quelle configuration est la plus efficace pour détecter un emploi sportif ou non. Seule la colonne 2 sera comparée aux résultats du programme. La première colonne permet simplement de se faire une première idée sur la corrélation entre contexte sportif et emploi sportif. Il s'agit d'un choix personnel, dans lequel le pourcentage ou le nombre de mots de LexSpo retrouvés dans le contexte de l'occurrence n'ont pas été pris en compte. Il permet de se faire une idée mais toute conclusion sur les résultats attendus serait un peu hâtive. Nous relevons toutefois certaines observations que nous décrivons dans le chapitre IV.

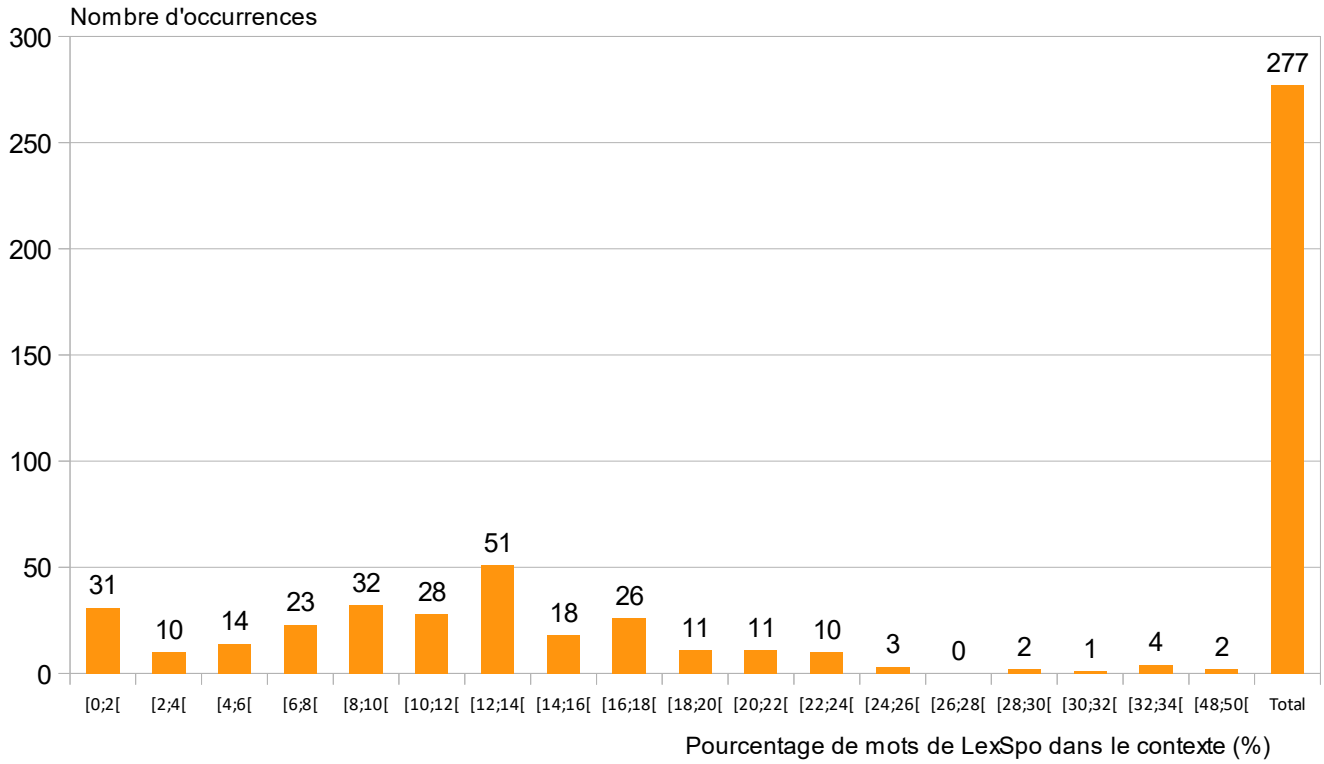
# CHAPITRE IV : VÉRIFICATION DE L'HYPOTHÈSE

Pour savoir si le contexte dans lequel un mot apparaît permet réellement de déduire automatiquement le type d'emploi de ce mot, nous devons ajuster certains paramètres du programme qui doivent nous permettre d'obtenir des résultats plus précis. Mais d'abord, nous cherchons à en savoir plus sur la relation entre contexte et emploi de spécialité en effectuant des analyses statistiques des fichiers annotés.

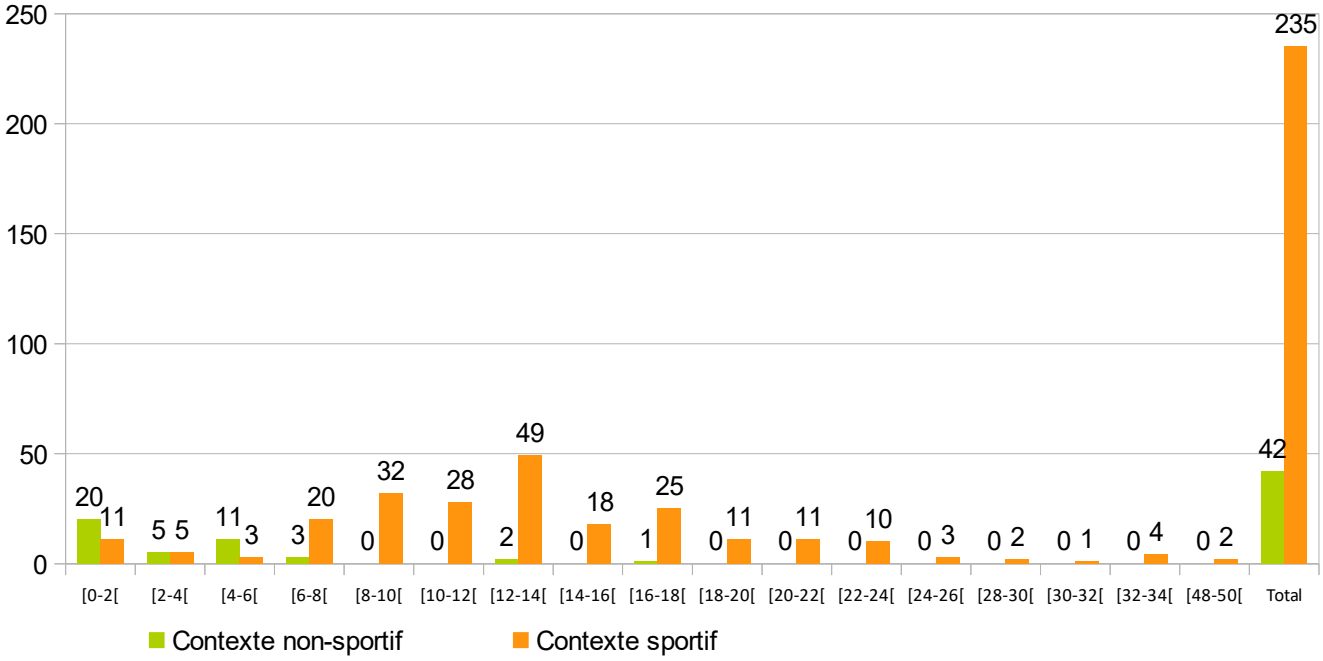
## I. OBSERVATIONS STATISTIQUES DES FICHIERS DE SORTIE POUR LA LISTE MONOSPORT

Les statistiques effectuées sur l'annotation manuelle des fichiers de sortie obtenus avec la liste monoSport montrent que l'ensemble des 277 occurrences extraites de LMglob correspondent à un emploi sportif du mot projeté. 84,8% de ces occurrences (soit 135) ont un contexte jugé sportif, contre 15,2% (soit 42) pour lesquelles ce n'est pas le cas. Ces statistiques mettent en évidence que les mots monosémiques marqués SPORT et autres domaines sportifs apparaissent majoritairement dans des contextes sportifs. Dans LMsport, les résultats révèlent que 96% des occurrences analysées (soit 145 sur 151) apparaissent dans un contexte sportif et 4% ont un contexte non sportif. Dans LMautre, il y a 126 occurrences et parmi celles-ci, 70,7% (soit 90 sur 126) ont un contexte sportif et 29,3% n'ont pas un contexte sportif. Ces statistiques confirment donc que les articles de sport contiennent des éléments de contexte qui favorisent la détection d'emplois sportifs. Cela signifie qu'en travaillant avec la liste polySport, le programme devrait détecter dans LMsport une plus grande proportion d'emplois sportifs que d'emplois non sportifs. En revanche, dans LMautre, il devrait y avoir une majorité d'emplois non sportifs. Si c'est le cas, alors il s'agira d'un élément allant dans le sens de l'hypothèse selon laquelle un mot de spécialité apparaît dans son contexte de spécialité.

Le graphique 1 présente la répartition de l'ensemble des occurrences extraites de LMGlob en fonction du pourcentage de mots de LexSpo trouvés dans leur contexte. Le graphique 2 présente les mêmes informations mais en dissociant les les contextes sportifs et non sportifs. Nous avons déterminé que 92,9% des 42 emplois annotés comme apparaissant dans un contexte non sportif contiennent entre 0 et 10% de mots appartenant à LexSpo dans leur contexte (soit 39 occurrences). Voici ce que nous pouvons dire de ces emplois pour expliquer que, malgré leur emploi relevant du domaine sportif, ils aient un faible poucentage de mots de LexSpo dans leur contexte.




*Graphique 1: répartition de l'ensemble des occurrences issues de LMglob des mots de monoSport en fonction du pourcentage de mots de LexSpo trouvés dans leur contexte.*




*Graphique 2: répartition des types de contextes issus de LMGlob pour les occurrences des mots de monoSport en fonction du pourcentage de mots de LexSpo qui s'y trouvent.*

Pour neuf d'entre eux, il s'agit de mots comme boxeur ou footballeur qui désignent des sportifs mais dans un contexte différent de leur métier. Il peut s'agir d'un article sur leur vie privée dans lequel on les désigne par « le fooballeur » ou « le boxeur ».

*« LE **FOOTBALLEUR** ARGENTIN DIEGO MARADONA, 39 ans, a été hospitalisé mardi 4 janvier dans une clinique privée de la station balnéaire de Punta del Este. » (tiré de l'article id="LM10-d683491p2" rub="SPO")*

Il peut également s'agir d'un article sur une personnalité sportive qui s'est reconvertie et dont on parle dans un tout autre contexte.

*« Bill Bradley vient de créer une autre surprise en ramassant, au cours du quatrième trimestre, presque autant d'argent […] que George Bush […], loin devant le sénateur McCain […] et Al Gore […]. Ce qui montre que l'ancienne star du **basket-ball** professionnel qu'est M. Bradley dissimule, derrière son image de politicien différent, d'étroites relations avec les milieux économiques et Wall Street. » (tiré de l'article id="LM10-d683431p2" rub="INT")*

Nous trouvons aussi des emplois de ces mots dans des récits et parfois pour qualifier des objets ou des comportements.

*« Un moine trouve un singe doté de la parole. Il s'enfuit avec lui et un **boxeur** traqué par des gangsters. Un singe qui parle plus un **footballeur** devenu acteur, ça ne suffit pas pour que la sauce prenne. » (tiré de l'article id="LM10-d686726p10" rub="TEL")*

*« Intelligent, bien bâti, physique de **boxeur** et tête bien faite, ce flic aigu et efficace n'hésite pas, quand l'occasion se présente, à faire lui-même justice. » (tiré de l'article id="LM10-d683806p4" rub="POC")*

Nous constatons que certains emplois s'insèrent dans un contexte politique, sociologique ou de reportage, lorsque le sport fait l'objet d'études et de débats au cours desquels il n'y a pas nécessairement d'utilisation de mots spécialisés.

*« Jean Tiberi (RPR), maire de Paris, a indiqué, vendredi 7 janvier, qu'il souhaite reconvertir l'un des trois **hippodromes** parisiens en espace de loisir destiné aux Parisiens. » (tiré de l'article id="LM10-d684158p2" rub="DER")*

*« Le champion de **foot** ou le héros de sitcom ne sert donc plus seulement à vendre de l'audience, mais à être le facteur de différenciation qui va attirer le client vers d'autres formes de consommation. » (tiré de l'article id="LM10-d685590p3" rub="HOR")*

Notons que le mot hippodrome apparaît systématiquement dans un contexte non sportif en rapport avec le débat sur la reconversion de l'infrastructure en espace de loisir. Nous n'observons pas d'occurrence apparaissant en contexte sportif pour ce mot-là, mais il n'apparaît que 9 fois dans le corpus et jamais dans les articles de sport. C'est trop peu pour généraliser et en conclure que certains mots du sport n'apparaissent jamais en contexte sportif et c'est également problématique si c'est le cas, parce que cela signifie qu'il faudrait mettre au point des méthodes complémentaires pour détecter le type d'emploi de ces mots là.

Enfin, nous relevons pour le mot match des emplois qui font référence au sens trouvé dans GLAWI, à savoir une « lutte entre deux concurrents ou deux équipes, rencontre (sportive). », sans pour autant être utilisés dans un contexte sportif.

> *« En octobre, il a même ajouté qu'Augusto Pinochet devait être jugé. Un geste qui lui a permis de gagner des électeurs, qui auraient pu être effrayés par son passé de fonctionnaire pinochétiste, et qui explique en partie l'inattendu* ***match*** *nul du premier tour des élections présidentielles le 12 décembre. » (tiré de l'article id="LM10-d684985p7" rub="INT")*

Nous avons vérifié dans le TLFi, le mot *match* y est également considéré comme monosémique et son sens est étiqueté SPORT. La différence avec GLAWI est qu'il y a une mention PAR MÉTAPHORE qui indique que cet emploi est aussi utilisé dans d'autres contextes que le sport et avec le même sens. Nous pouvons parler d'une forme de despécialisation ici, ce qui explique pourquoi le pourcentage de mots de LexSpo dans le contexte de cette occurrence est faible (2,54%).

Dans tous les cas, très peu de mots de contexte appartenant à LexSpo apparaissent autour de ces emplois. Ceux dont le pourcentage est supérieur à 10% (3 occurrences) sont employés dans des paragraphes d'articles où il n'y a pas ou peu de mots du sport. Le propos principal est centré sur un autre sujet mais les quelques mots du sport suffisent à faire monter les pourcentages. C'est positif car il s'agit bien d'un emploi sportif que nous avions considéré comme apparaissant dans un contexte non sportif, ce qui signifie que selon le seuil, ce type d'emploi pourrait être détecté.

> *« Alors que les auditions se poursuivent dans l'affaire des comptes du club de basket- ball du Cercle Saint-Pierre de Limoges, l'équipe, engagée en Coupe Korac, devait se rendre à Kiev (Ukraine), mercredi 20 janvier, pour gagner sa place en quart de finale d'une compétition qu'elle a déjà emportée à deux reprises et retrouver un peu de son lustre et de sa confiance. Le président du CSP, Jean-Paul de Peretti, entendu par les enquêteurs et mis en examen, comme cinq autres responsables du club, se démène pour monter un plan de sauvetage et convaincre un partenaire privé d'investir afin de « gommer » un déficit estimé au moins à 5 millions de francs. Paradoxalement, les candidats sur les rangs avaient été sollicités par Didier Rose, dirigeant limougeaud influent, richissime agent de joueurs, homme fort du* ***basket-***

*__ball__ français, mis en examen et incarcéré le 12 janvier. » (tiré de l'article id="LM10-d685865p1" rub="SPO")*

Pour les occurrences annotées comme apparaissant dans un contexte sportif, nous en dénombrons 70 qui ont moins de 10% de mots spécialisés dans leur contexte (soit 29,9% de l'ensemble des occurrences) et 164 qui ont plus de 10% de mots spécialisés dans leur contexte (soit 70,1%).

Nous concluons pour l'analyse des annotations de la liste monoSport en rappelant que dans presque 85% des cas, un mot issu du domaine sportif apparaît dans un contexte également composé de mots du sport. Il s'agit maintenant de savoir comment se manifestent les emplois non sportifs. En effet, si ces derniers apparaissent dans un contexte composé de mots spécialisés également, la démarche n'est pas pertinente. Si ce n'est pas le cas, nous pouvons continuer et tenter de déterminer une valeur de pourcentage seuil en deçà de laquelle nous considérons chaque emploi comme non sportif.

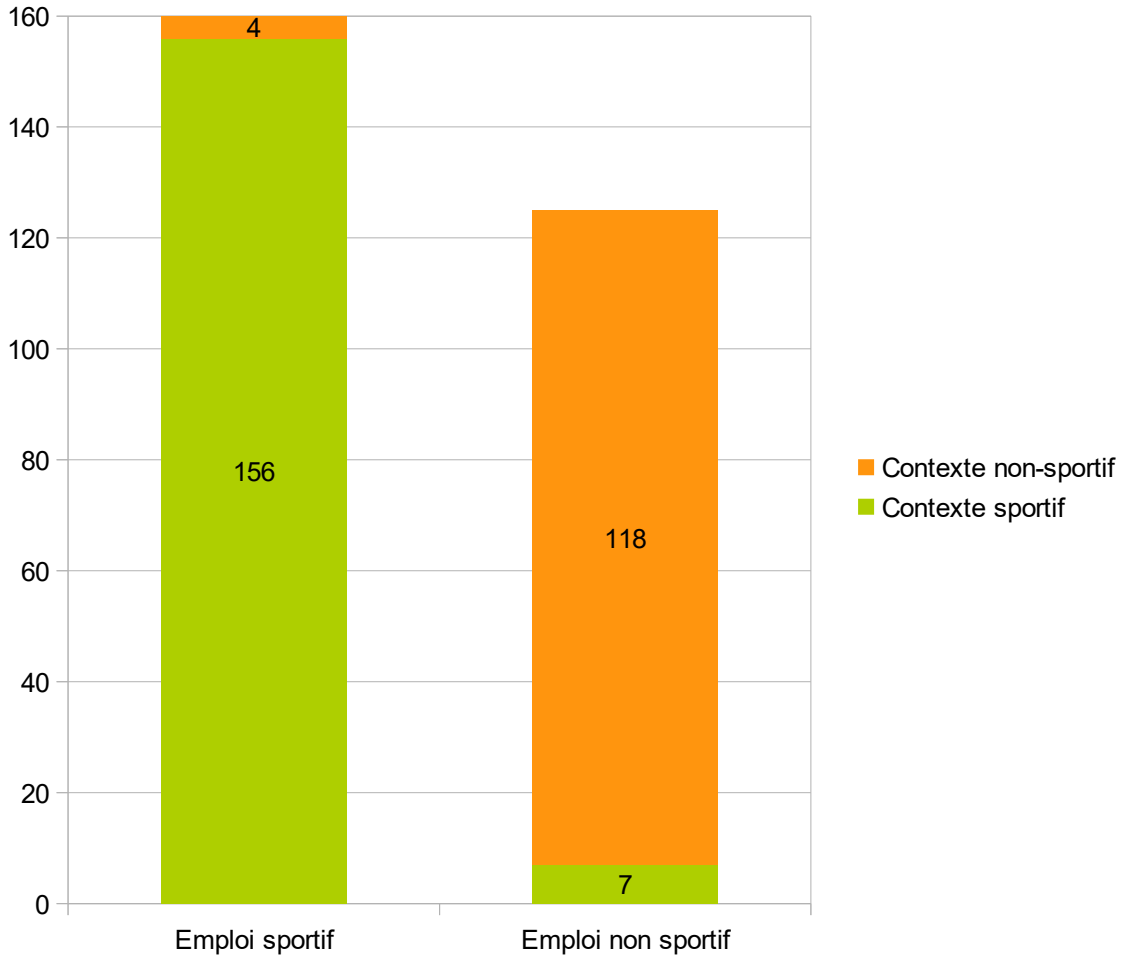
## II. OBSERVATIONS STATISTIQUES DES FICHIERS DE SORTIE POUR LA LISTE POLYSPORT

Observons à présent les résultats de l'annotation manuelle pour les mots issus de polySport. Nous cherchons à savoir si, pour des mots indiqués polysémiques dans le dictionnaire, il serait possible de savoir automatiquement si nous avons affaire à un emploi spécialisé ou non. Les résultats précédents nous laissent entendre que la majorité des emplois spécialisés ont un contexte composé de mots spécialisés. Il nous reste à savoir si les emplois non spécialisés sont utilisés dans un contexte non spécialisé.

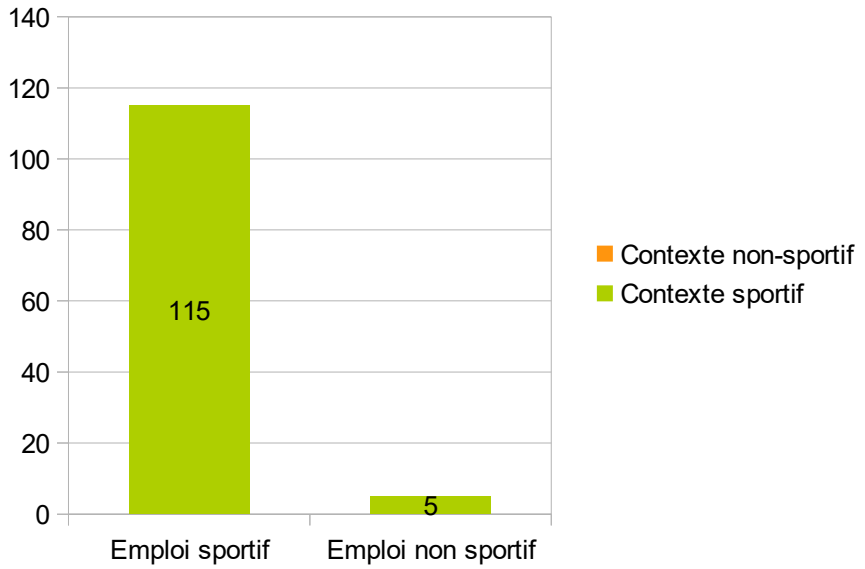
Selon l'annotation, parmi les 285 acceptions extraites de LMglob, 125 (44,2%) ont un emploi non-spécialisé dans le domaine du sport, et 160 (55,8%) ont un emploi spécialisé. Dans le sous-corpus LMsport, toutes les occurrences extraites apparaissent dans un contexte sportif, mais sur les 120, cinq d'entre elles ne relèvent pas d'un emploi sportif, ce qui correspond à 4,2% du sous-corpus. Les 95,8% restants relèvent bien d'un emploi sportif. Dans le sous-corpus LMautre, il y a 165 occurrences. Parmi les emplois non sportifs de LMautre (qui représentent 72,73% du sous-corpus), 1,67% ont un contexte sportif et 98,33% ont un contexte non sportif. Parmi les emplois sportifs, 8,89% ont un contexte non sportif et 91,11% ont un contexte sportif. Dans la grande majorité des cas, les emplois sportifs ont un contexte sportif et les emplois non sportifs ont un contexte non sportif. Le graphique 3 schématise la répartition des contextes sportifs et non sportifs selon le type d'emploi.

| | | Emplois sportifs | Emplois non sportifs | Total |
|---|---|---|---|---|
| **LMsport** | **Contextes sportifs** | 115 | 5 | 120 |
| | **Contextes non sportifs** | 0 | 0 | 0 |
| | **Total** | 115 | 5 | 120 |
| **LMautre** | **Contextes sportifs** | 41 | 2 | 43 |
| | **Contextes non sportifs** | 4 | 118 | 122 |
| | **Total** | 45 | 120 | 165 |
| **LMglob** | **Contextes sportifs** | 156 | 7 | 163 |
| | **Contextes non sportifs** | 4 | 118 | 122 |
| | **Total** | 160 | 125 | 285 |

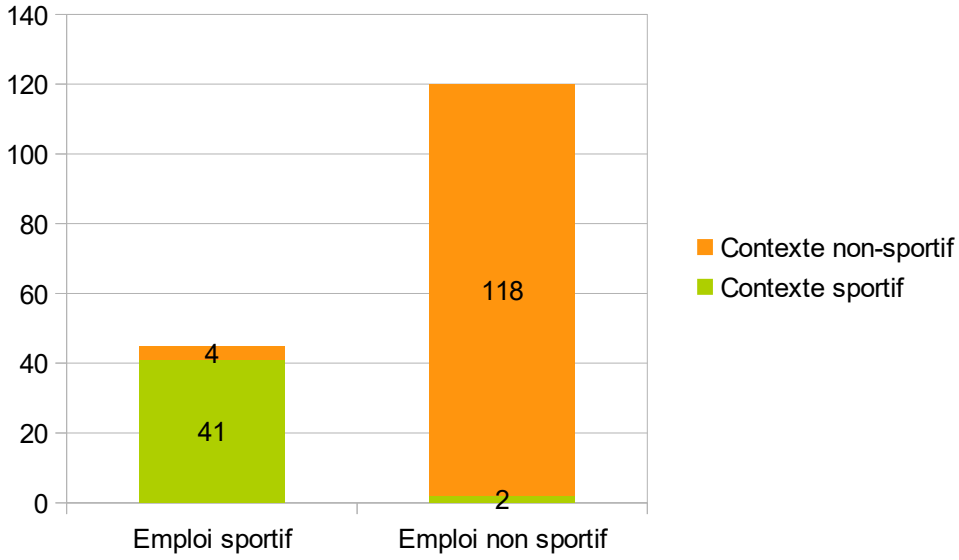
*Tableau 10 : synthèse des résultats obtenus avec la liste polySport.*




*Graphique 3: répartition des contextes sportifs et non sportifs en fonction du type d'emploi dans LMglob.*




*Graphique 4: répartition des contextes sportifs et non sportifs en fonction du type d'emploi dans LMsport.*

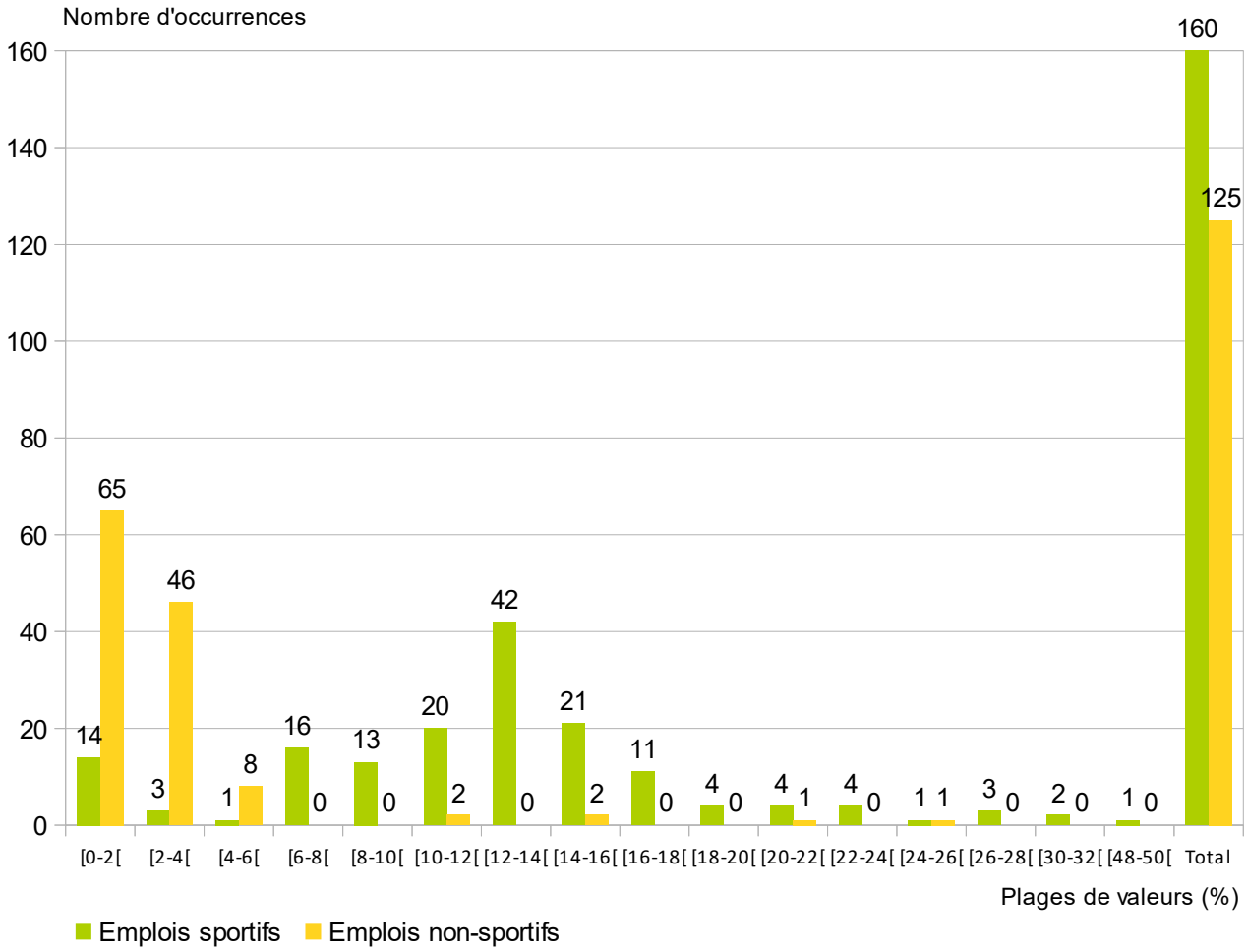



*Graphique 5: répartition des contextes sportifs et non sportifs en fonction du type d'emploi dans LMautre.*

Nous pouvons déduire de ces observations qu'il existe bien un lien entre le type d'emploi et la nature du contexte dans la majorité des cas. Il convient néanmoins de nuancer l'interprétation

de ces résultats en rappelant qu'il s'agit d'observations effectuées sur une liste de mots succincte et un petit échantillon de données dont nous ne pouvons pas affirmer la représentativité. Des études sur plus de mots et un corpus plus large seraient nécessaires pour avoir des résultats s'approchant au maximum de la réalité. Ceci étant dit, les statistiques permettent tout de même de relever les tendances que nous venons de décrire. Ces tendances corroborent notre hypothèse et nous permettent de poursuivre nos manipulations.

Le graphique 6 présente le nombre d'occurrences sportives et non-sportives réparties selon le pourcentage de mots appartenant à LexSpo dans leur contexte. Il permet de visualiser comment se répartissent les différents types d'emplois et quelles tendances se dégagent de l'aspect du contexte en fonction du type d'emploi observé.




*Graphique 6: répartition des occurrences sportives et non sportives de LMglob en fonction du pourcentage de mots de LexSpo dans leur contexte.*

Nous remarquons que le contexte de la plupart des emplois non sportifs est pauvre en mots appartenant à LexSpo. En revanche, 88,75% des emplois sportifs ont un contexte composé

d'au moins 6% de mots de LexSpo. Nous voyons bien ici que le type d'emploi a un lien avec la nature spécialisée ou non du contexte dans lequel il apparaît. Nous partons de ce graphique pour fixer une valeur seuil initiale et l'ajusterons par la suite. Le graphique 6 indique qu'à partir de 6% il y a une inversion de la tendance. Entre 0 et 6, le nombre d'emplois sportif est inférieur au nombre d'emplois non sportifs. À partir de 6%, les emplois sportifs deviennent plus nombreux que les emplois non sportifs. Cette tendance s'explique en partie par le fait que le nombre total d'emplois non sportifs est largement inférieur au nombre total d'emplois sportifs, mais nous constatons que 119 occurrences ayant des emplois non sportifs ont entre 0% et 6% de mots de LexSpo dans leur contexte. Cela représente 95,2% des emplois non sportifs. D'autre part, les occurrences ayant des emplois sportifs et dont le pourcentage de mots de LexSpo dans leur contexte est inférieur à 6 sont au nombre de 13, soit 11,25% de l'ensemble des emplois sportifs. En fixant la valeur seuil à 6%, toutes les occurrences ayant dans leur fenêtre contextuelle plus de 6% de mots de LexSpo sont considérées comme des emplois sportifs. Le tableau 11 présente les scores que nous obtenons avec ce seuil.

| s=6 | Emplois sportifs | Emplois non sportifs | Total |
|---|---|---|---|
| Précision | 0,96 | 0,87 | 0,91 |
| Rappel | 0,89 | 0,95 | 0,92 |
| F-mesure | 0,92 | 0,91 | 0,92 |

*Tableau 11: scores de précision, rappel et f-mesure pour une valeur seuil de 6%*

Ces scores laissent penser que dans le domaine du sport, le lien entre un terme et son contexte est suffisant, dans la plupart des cas, pour qu'on puisse considérer qu'il y a une démarcation nette entre la proportion de mots spécialisés dans le contexte d'un emploi sportif et d'un emploi non sportif. Examinons à présent dans quelle mesure il est possible d'ajuster la valeur seuil pour obtenir de meilleurs résultats.

# III. AJUSTEMENT DE LA VALEUR SEUIL

## 1. Objectif et procédure

Notre objectif est de trouver une valeur seuil qui nous rapproche au plus près des résultats obtenus avec l'annotation manuelle. Nous lançons le programme sur le corpus entier avec la liste monoSport, en changeant la valeur du seuil à chaque *run*. Nous obtenons en sortie un fichier tabulé semblable à celui du fichier d'annotation mais dans lequel la dernière colonne correspond à

la réponse du programme selon que le pourcentage obtenu est supérieur ou inférieur à la valeur seuil. Le tableau 12 illustre le format du fichier de sortie obtenu.

| TERME PROJETÉ | NOMBRE DE MOTS DU LEXIQUE | % DE MOTS DU LEXIQUE | OCCURRENCE EN CONTEXTE. | GOLD | pourcentage>s (où s=6) (0/1) |
|---|---|---|---|---|---|
| arbitre | 77 | 6,32% | […] Humiliés par la prestation de leurs **joueurs** menés 3 à 0 au bout de vingt minutes, le millier de supporteurs de l'**OM** ayant fait le déplacement manifestent leur dépit en lançant des fumigènes et en arrachant des sièges. L'**arbitre** Pascal Garibian est obligé d'interrompre la rencontre pendant une dizaine de minutes alors qu'une partie des fans marseillais préfèrent quitter le **stade**.[...] | 1 | 1 |
| challenger | 1 | 25,00% | L'avancée des **challengers** dans la **course** à la Maison Blanche | 0 | 1 |
| marathon | 5 | 1,45% | […] Au terme d'un épuisant **marathon**, les négociateurs sont parvenus à boucler lundi un pacte qui laisse sur la touche la formation de Jörg Haider, le FPÖ - créditée désormais de 30 % des voix dans les sondages, contre 27 % lors du scrutin du 3 octobre -, mais prévoit des coupes budgétaires forcément impopulaires, et de nouvelles privatisations du secteur public. [...] | 0 | 0 |
| médaille | 47 | 13,24% | […] Au début du mois de novembre 1999, aux **championnats** du monde, à Séoul (Corée du Sud), les **épéistes**, **fleurettistes** et **sabreurs** ont empoché huit **médailles**, dont cinq d'or et une d'argent, en individuel comme en **équipe**.En attendant son abri provisoire, l'**escrime** française se console en espérant une reconstruction, et donc une modernisation de locaux qui devenaient un peu vétustes. [...] | 1 | 1 |
| peloton | 0 | 0,00% | ÉCART CREUSÉ AVEC LE PELOTON | 1 | 0 |
| basket-ball | 18 | 25,00% | En raison d'un manque de vent, les **régates** des **demi-finales** de la **Coupe** Louis-Vuitton, **qualificative** à la **Coupe** de l'America, prévues pour jeudi 6 janvier ont été annulées. En outre, le **défi** américain Stars-&-Stripes a obtenu un délai supplémentaire de 24 heures pour réparer son **voilier** endommagé lors d'une collision, le 4 janvier, avec 6e Sens, qui a valu 0,5 point de **pénalité** au **bateau** du **défi** français.[...] | 1 | 1 |
| pompe | 11 | 5,42% | « La situation était stabilisée. »Elle se dégrade à nouveau à partir de minuit. D'abord par une nouvelle gîte à tribord. Les **pompes** à eau sont relancées dans les ballasts. « L'état de la mer se détériorait et le navire embarquait, précise le capitaine [...] » | 0 | 0 |
| saut | 3 | 30,00% | **SKI** NORDIQUE : l'Allemand Martin Schmitt **remporte** les deux concours de **saut** à **skis** de Sapporo. | 1 | 1 |
| stade | 50 | 9,67% | Les spectateurs du **stade** Furiani n'ont pas rêvé. Ils ont bien vu, mercredi 12 janvier, trois ou quatre **joueurs** de l'**Olympique** de Marseille lever les bras au ciel, comme en guise de **victoire**, à l'issue d'une rencontre sans **but**. [...] | 1 | 1 |

*Tableau 12: exemple de format de sortie obtenu en sortie du programme.*

Au vu du graphique 6, la valeur seuil la plus pertinente se situe autour de 6%, car c'est à partir de cette valeur que le nombre d'occurrences sportives devient supérieur au nombre d'occurrences non sportives. Nous décidons donc d'étudier les fichiers de sortie pour des seuils compris entre 4% et 8%, avec des intervalles de 0,25% à chaque run. Nous lançons le programme avec la liste polySport car nous voulons des résultats pour des emplois non sportifs. Observer le comportement du programme lorsqu'il doit analyser des contextes pour des emplois non sportifs permet de se faire une idée des résultats que nous pourrions obtenir sur un cas de déterminologisation. D'autre part, tenter de fixer un seuil à partir de la liste monoSport ne serait pas pertinent. En effet, puisque la totalité des occurrences extraites relèvent d'emplois sportifs, le meilleur score que nous pourrions obtenir serait 100, en fixant le seuil à 0. Mais nous n'aurions alors aucun moyen de savoir quelle serait la précision et le rappel pour des emplois non sportifs, possiblement déterminologisés.

## 2. Résultats

Les tableaux 13 et 14 présentent le détail des résultats obtenus avec le programme pour des seuils compris entre 4% et 8% avec la liste PolySport projetée sur les 285 occurrences du corpus LMglob.

| **Nb emplois sportifs (gold) : 160** | **Emplois sportifs** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Seuils (S)** | **Nb emplois sportifs** | **Nb emplois non sportifs** | **VP** | **FP** | **FN** | **VN** | **Précision** | **Rappel** | **F-mesure** |
| **S=4,00** | 157 | 128 | 143 | 14 | 17 | 111 | 0,911 | 0,894 | 0,902 |
| **S=4,25** | 156 | 129 | 143 | 13 | 17 | 112 | 0,917 | 0,894 | 0,905 |
| **S=4,50** | 153 | 132 | 143 | 10 | 17 | 115 | 0,935 | 0,894 | 0,914 |
| **S=4,75** | 153 | 132 | 143 | 10 | 17 | 115 | 0,935 | 0,894 | 0,914 |
| **S=5,00** | 153 | 132 | 143 | 10 | 17 | 115 | 0,935 | 0,894 | 0,914 |
| **S=5,25** | 150 | 135 | 142 | 8 | 18 | 117 | 0,947 | 0,888 | 0,916 |
| **S=5,50** | 149 | 136 | 142 | 7 | 18 | 118 | 0,953 | 0,888 | 0,919 |
| **S=5,75** | 148 | 137 | 142 | 6 | 18 | 119 | 0,959 | 0,888 | 0,922 |
| **S=6,00** | 148 | 137 | 142 | 6 | 18 | 119 | 0,959 | 0,888 | 0,922 |
| **S=6,25** | 147 | 138 | 141 | 6 | 19 | 119 | 0,959 | 0,881 | 0,919 |
| **S=6,50** | 141 | 144 | 135 | 6 | 25 | 119 | 0,957 | 0,844 | 0,897 |
| **S=6,75** | 139 | 146 | 133 | 6 | 27 | 119 | 0,957 | 0,831 | 0,890 |
| **S=7,00** | 139 | 146 | 133 | 6 | 27 | 119 | 0,957 | 0,831 | 0,890 |
| **S=7,25** | 139 | 146 | 133 | 6 | 27 | 119 | 0,957 | 0,831 | 0,890 |
| **S=7,50** | 136 | 149 | 130 | 6 | 30 | 119 | 0,956 | 0,813 | 0,878 |
| **S=7,75** | 133 | 152 | 127 | 6 | 33 | 119 | 0,955 | 0,794 | 0,867 |
| **S=8,00** | 132 | 153 | 126 | 6 | 34 | 119 | 0,955 | 0,788 | 0,863 |

*Tableau 13 : résultats obtenus pour les emplois sportifs des mots de la liste polySport pour des seuils de 4,00 à 8,00 sur le corpus LMglob.*

Nous voyons ici que pour les emplois sportifs, la précision augmente en même temps que le seuil, culmine pour des valeurs de seuil comprises entre 5,75 et 6,25 et diminue de nouveau. Nous expliquons ce phénomène ainsi : la précision correspond au rapport entre les vrais positifs et le nombre total d'emplois considérés comme sportifs (la somme entre les vrais positifs et les faux positifs). Or, nous pouvons observer que le nombre de faux positifs (c'est-à-dire les emplois non sportifs dont le pourcentage est supérieur au seuil) ne bouge plus à partir de 5,75%. Ceci est dû au fait que les six emplois non-sportifs restants ont un pourcentage supérieur ou égal à 10%. En augmentant le seuil, nous perdons donc en précision parce que le nombre de vrais positifs diminue alors que le nombre de faux positifs reste au même stade. Pour que le nombre de faux positifs diminue, et donc que la précision augmente, il faudrait fixer le seuil plus haut. Cependant nous perdrions en rappel, qui lui, diminue en même temps que la valeur de seuil augmente. Le meilleur rapport entre les scores de précision et de rappel est obtenu pour des valeurs seuil de 5,75% et 6%. Cela signifie que c'est avec ces valeurs de seuil que nous détectons le plus efficacement les emplois sportifs. Observons maintenant dans les résultats obtenus pour les emplois non sportifs.

| Nb emplois non sportifs (gold) : 125 | Emplois non sportifs | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Seuils (S)** | **Nb emplois sportifs** | **Nb emplois non sportifs** | **VP** | **FP** | **FN** | **VN** | **Précision** | **Rappel** | **F-mesure** |
| **S=4,00** | 157 | 128 | 111 | 17 | 14 | 143 | 0,867 | 0,888 | 0,877 |
| **S=4,25** | 156 | 129 | 112 | 17 | 13 | 143 | 0,868 | 0,896 | 0,882 |
| **S=4,50** | 153 | 132 | 115 | 17 | 10 | 143 | 0,871 | 0,920 | 0,895 |
| **S=4,75** | 153 | 132 | 115 | 17 | 10 | 143 | 0,871 | 0,920 | 0,895 |
| **S=5,00** | 153 | 132 | 115 | 17 | 10 | 143 | 0,871 | 0,920 | 0,895 |
| **S=5,25** | 150 | 135 | 117 | 18 | 8 | 142 | 0,867 | 0,936 | 0,900 |
| **S=5,50** | 149 | 136 | 118 | 18 | 7 | 142 | 0,868 | 0,944 | 0,904 |
| **S=5,75** | 148 | 137 | 119 | 18 | 6 | 142 | 0,869 | 0,952 | 0,908 |
| **S=6,00** | 148 | 137 | 119 | 18 | 6 | 142 | 0,869 | 0,952 | 0,908 |
| **S=6,25** | 147 | 138 | 119 | 19 | 6 | 141 | 0,862 | 0,952 | 0,905 |
| **S=6,50** | 141 | 144 | 119 | 25 | 6 | 135 | 0,826 | 0,952 | 0,885 |
| **S=6,75** | 139 | 146 | 119 | 27 | 6 | 133 | 0,815 | 0,952 | 0,878 |
| **S=7,00** | 139 | 146 | 119 | 27 | 6 | 133 | 0,815 | 0,952 | 0,878 |
| **S=7,25** | 139 | 146 | 119 | 27 | 6 | 133 | 0,815 | 0,952 | 0,878 |
| **S=7,50** | 136 | 149 | 119 | 30 | 6 | 130 | 0,799 | 0,952 | 0,867 |
| **S=7,75** | 133 | 152 | 119 | 33 | 6 | 127 | 0,783 | 0,952 | 0,859 |
| **S=8,00** | 132 | 153 | 119 | 34 | 6 | 126 | 0,778 | 0,952 | 0,856 |

*Tableau 14 : résultats obtenus pour les emplois non sportifs des mots de la liste polySport pour des seuils de 4,00 à 8,00 sur le corpus LMglob.*

Ici, la tendance est inversée. Plus le seuil augmente, plus les emplois non-sportifs sont correctement considérés comme non sportifs et meilleur est le rappel. En revanche, la précision diminue lorsque le seuil augmente car le bruit est plus important (des emplois sportifs sont considérés comme non-sportifs). Pour la précision, nous observons de nouveau un phénomène en parabole, où le score augmente pour diminuer ensuite. Il s'explique par le fait que le nombre de vrais positifs (emplois non sportifs considérés à raison comme emplois non sportifs) augmente rapidement puis stagne à partir de 5,75% tandis que le nombre de faux positifs (emplois sportifs considérés à tort comme emplois non sportifs) continue d'augmenter à mesure que le seuil augmente. Le meilleur score de précision serait obtenu pour une valeur de seuil basse, car une majorité d'emplois sportifs seraient considérés comme tels, mais le rappel serait mauvais parce que beaucoup d'emplois non sportifs ne seraient pas ramenés. Ici, la meilleure F-mesure est obtenue pour des valeurs seuil de 5,75% et 6%.

Pour ce travail notre objectif principal est de détecter correctement des emplois non sportifs. Nous cherchons donc à avoir la meilleure précision possible pour les emplois non sportifs, afin que le lexicographe n'ait pas à traiter des données non-pertinentes qui lui font perdre du temps, mais nous voulons également un rappel correct car plus les données sont fournies, plus l'analyse du lexicographe est précise. Nous obtenons donc les meilleurs résultats en fixant la valeur seuil à 5,75% ou 6%. Nous décidons de garder le seuil minimum pour que, sur un autre corpus ou avec une autre liste, le rappel soit potentiellement plus précis qu'à 6%. Nous effectuons les manipulations suivantes avec un seuil à 5,75%.

Les résultats obtenus confirment bien qu'en milieu polysémique, il y a un lien entre le contexte et l'emploi. La plupart du temps, lorsque l'occurrence relève d'un emploi sportif, elle se retrouve entourée d'un contexte sportif, en revanche, lorsque l'occurrence relève d'un emploi non sportif, elle se retrouve dans un contexte non sportif.

## 3. Analyse des erreurs

Nous distinguons trois catégories principales d'erreurs. La première regroupe 17 occurrences, soit 70,8% des erreurs. Il s'agit du cas de figure où les occurrences relèvent d'un emploi sportif mais n'apparaissent pas dans un contexte suffisamment pourvu en mots du sport pour être détectés. Le pourcentage n'atteint pas le seuil et ces occurrences sont donc déclarées comme non-sportives. Voici par exemple une occurrence du mot *attaquant*, qui a 1,7% de mots du sport dans son contexte :

> *« Amenés sur des palanquins, la chanteuse Céline Dion et son mari René Angélil ont renouvelé leurs voeux de mariage, mercredi 5 janvier, à Las Vegas, en présence d'oiseaux exotiques et de chameaux. Les parents de Tuatahi Manaakitunga, « la première bénédiction » en Maori , premier bébé de l'an 2000, qui, d'après leur*

*agent, Andy Haden, ancien **attaquant** des All Blacks, accorderaient volontiers leur exclusivité à un magazine international, s'alarment au chevet du nouveau-né, opéré du coeur à Auckland, jeudi 6 janvier. Mardi 4 janvier, au service des urgences du Northwick Park Hospital, établissement public londonien manquant de lits et de personnel, débordant de malades allongés sur des brancards, les derniers admis étaient soignés dans des camionnettes. [...] » (tiré de l'article id="LM10-d684020p2", rub="COM")*

Le mot *attaquant* est bien employé dans son sens sportif, mais il apparaît dans un contexte non sportif.

Le deuxième type d'erreur concerne une occurrence, soit 4,2% des erreurs, et correspond au cas où les articles, donc les fenêtres contextuelles, sont très courtes. L'occurrence a un emploi et un contexte non sportif mais il suffit d'un mot appartenant à LexSpo pour faire monter le pourcentage de manière importante. En voici un exemple :

*« L'avancée des **challengers** dans la course à la Maison Blanche. » (tiré de l'article id="LM10-d683431p1" rub="INT")*

L'article se résume à cette seule phrase, donc la fenêtre contextuelle également. Il y a un mot dans cette fenêtre qui fait partie de LexSpo : *course*. En tout, quatre mots sont pris en compte dans le calcul de pourcentage : *avancée*, *course*, *Maison* et *Blanche*. Le score obtenu est donc de 25%.

Le troisième type d'erreur, qui représente 20,8% du nombre total d'erreurs, a lieu lorsqu'une occurrence relève d'un emploi non sportif mais apparaît dans un contexte sportif qui fait monter le pourcentage. Par exemple, pour cette occurrence du mot *stade* :

*« Compacte et très sûre d'elle, Patty Schnyder enlève la **première** manche à force de coups droits décalés et de services très travaillés. Elle se détache dans la **deuxième** manche, profitant d'un passage à vide de son **adversaire** au service. « J'ai tout simplement fait un grand **match**, cela faisait longtemps que cela ne m'était pas arrivé », a-t-elle commenté.Ironie du sort, c'est en **battant** celle-ci alors tête de **série** no 11 au même **stade** de la **compétition**, en 1999, qu'Amélie Mauresmo s'était révélée au public australien avant de s'avancer jusqu'en **finale** de l'**épreuve**. » (tiré de l'article id="LM10-d686052p3" rub="SPO")*

Nous détaillons dans le chapitre V (cf. p.59) les pistes envisagées pour améliorer les résultats.

## IV. VÉRIFICATION AVEC LA LISTE MONOSPORT

Nous avons montré que les emplois sportifs apparaissent majoritairement dans un contexte sportif et les emplois non sportifs dans un contexte non sportif. Cependant des ambiguïtés existent et nous souhaitons savoir quels résultats obtient le programme dans la détection d'emplois non ambigus. L'objectif est ici de voir quels résultats nous obtenons en lançant le programme avec la liste monoSport et avec le seuil fixé dans la partie précédente. Nous voulons un fichier de sortie avec le même format que celui présenté au tableau 12, ce qui nous permettra de comparer les résultats du programme avec les résultats de l'annotation. Le tableau 15 présente les scores obtenus avec la liste monosémique.

| | **Nb emplois sportifs** | **Nb emplois non sportifs** | **VP** | **FP** | **FN** | **VN** | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|---|---|---|---|---|---|
| **Emplois sportifs** | 222 | 55 | 222 | 0 | 55 | 0 | 1 | 0,8 | 0,89 |

*Tableau 15 : scores obtenus pour les mots de monoSport pour un seuil S=5,75.*

Les résultats obtenus sont inférieurs à ce que nous obtenons avec la liste polySport. La précision est de 1 parce qu'il n'y a aucun emploi annoté manuellement comme un emploi non sportif, toutes les occurrences dont le pourcentage de mots sportifs dans le contexte est supérieur à 5,75 sont donc à raison considérées comme relevant d'emplois sportifs. En revanche, le rappel est moins bon, car 36 occurrences ont un pourcentage inférieur à 5,75, ce qui représente 13% des occurrences extraites. Les erreurs sont les mêmes que celles décrites dans la partie I. de ce chapitre (cf. p.41).

Au vu de ces résultats et des améliorations possibles, nous pouvons dire que le contexte est une bonne solution pour détecter un emploi. Cependant, la langue pouvant être ambiguë, il arrive fréquemment qu'un terme soit utilisé dans un contexte non spécialisé, en conservant quand même son sens et son emploi d'origine. Des pistes pour améliorer les résultats sont proposées dans le chapitre V (cf. p. 59).

## V. VÉRIFICATION AVEC UNE LISTE POLYSÉMIQUE DIFFÉRENTE

Nous vérifions nos résultats avec des mots-cibles différents, afin de nous assurer que les résultats obtenus avec polySport ne sont pas uniquement dus au hasard ou au fait que nous avons choisi la valeur seuil pour laquelle les scores étaient les meilleurs. Nous établissons donc une troisième liste de mots-cibles construite sur le même principe que la liste polySport (cf chapitre III,

partie II. 1. b. p. 35). Nous choisissons cinq nouveaux termes parmi les mots polysémiques de GLAWI, différents de ceux de la liste polySport initiale. Cette fois, nous choisissons des mots exclusivement polysémiques, sans homonymes comme *stade*. Une déterminologisation n'entraîne pas nécessairement un changement radical de sens, il s'agit simplement d'un nouveau type d'emploi qui peut faire un peu évoluer le sens initial. Initialement, il s'agissait surtout de savoir si contexte et emploi étaient liés, les mots dont les différents sens étaient très différents étaient donc intéressants car ils offraient la perspective d'apparaître dans des contextes très variés. À présent, nous souhaitons voir si le type de contexte dans lequel apparaissent des mots polysémiques dont le sens est similaire d'un emploi à l'autre est aussi marqué. Nous sélectionnons les mots *coach*, *disputer*, *relais*, *revers* et *tapis*. Ils forment les éléments de la liste polySportBis.

Le mot *coach* est un anglicisme qui est passé dans le langage courant en français pour désigner un entraîneur sportif, mais également une figure de guide pour de nombreuses autres activités. Ces derniers temps, nous entendons fréquemment l'expression *coach en développement personnel* par exemple. *Coach* n'est donc plus cantonné au domaine sportif, et nous voulons savoir s'il est possible de déduire son type d'emploi automatiquement grâce à son contexte.

Dans le domaine sportif, *disputer* s'utilise dans l'expression *disputer un match*, c'est-à-dire lorsque deux équipes ou deux sportifs s'affrontent pour décrocher la victoire. Dans le langage courant, *disputer* fait également référence à un affrontement au cours duquel chacune des parties tente de l'emporter sur l'autre.

*Relais* est un mot qui désigne un type de course (à pied, dans l'eau,...) au cours de laquelle plusieurs sportifs forment une équipe et courent les uns après les autres, chacun attendant que le précédent ait terminé pour partir. Il désigne également l'objet, une sorte de manche en bois ou en plastique, qu'un coureur de relais doit passer au suivant de son équipe à l'arrivée pour prouver que ce dernier a bien attendu que celui qui le précède arrive pour partir. Dans le langage courant, *relais* est employé pour désigner une étape entre un point A et un point B ou une personne intermédiaire entre deux autres personnes.

En sport, un *revers* est un coup effectué généralement lorsque la balle ou le volant arrive du côté opposé à la main qui tient la raquette. Plus généralement, le revers désigne l'envers de quelque chose, la face cachée. L'emploi de ce mot peut s'avérer ambigu à déterminer car il apparaît souvent dans l'expression *le revers de la médaill*e. *Médaille* apparaissant dans LexSpo, l'expression peut être considérée comme ayant une nature sportive alors que ce n'est pas le cas.

Et enfin un *tapis* en sport est la surface sur laquelle se pratiquent certains sports de combat et la gymnastique. Dans le contexte général, il fait référence à une surface de tissu décorative que l'on met sur le sol. *Envoyer quelqu'un au tapis* est une expression qui provient du sport et qui s'est popularisée, elle signifie battre quelqu'un dans une compétition ou un concours, et est employée au sens propre, mais aussi au sens figuré pour toutes sortes de défis.

Nous procédons également à une phase d'annotation manuelle comme décrite au chapitre III, partie III. p.38), afin de comparer les résultats obtenus informatiquement à un étalon. Il y a 177 occurrences en tout. Nous déterminons 56 emplois sportifs et 121 emplois non-sportifs. Le tableau 16 présente les scores obtenus avec la liste polySportBis sur le corpus LMglob.

| | Nb emplois sportifs | Nb emplois non sportifs | VP | FP | FN | VN | Précision | Rappel | F-mesure |
|---|---|---|---|---|---|---|---|---|---|
| **Emplois sportifs** | 58 | 119 | 51 | 7 | 5 | 114 | 0,879 | 0,911 | 0,895 |
| **Emplois non sportifs** | | | 114 | 5 | 7 | 51 | 0,958 | 0,942 | 0,950 |

*Tableau 16: scores obtenus pour les mots de la liste polySportBis sur le corpus LMglob.*

Les résultats sont semblables à ceux obtenus avec la liste polySport, ce qui permet d'écarter la possibilité qu'ils soient dus au hasard ou au fait que nous ayons choisi le seuil qui donnait les meilleurs résultats.

## VI. CONCLUSION DES MANIPULATIONS

La conclusion de cette série d'analyses et de manipulations est que les emplois sportifs et non sportifs se distinguent bien, la plupart du temps, grâce au contexte dans lequel ils apparaissent. Cependant le contexte peut être créé par quelques mots seulement, aussi le calcul de pourcentage n'est peut-être pas l'outil le plus pertinent pour trancher sur la nature sportive ou non d'un emploi. Parfois, l'emploi sportif peut apparaître dans une expression au milieu d'un contexte non sportif, les mots qui créent son contexte peuvent être seulement au nombre de 2. Par exemple, voici une occurrence prélevée dans le fichier de sortie du dernier *run* analysé :

> *« A la veille du scrutin législatif, son service de presse pris cependant soin d'apporter une dernière touche à son image. Vladimir Poutine dans le rôle du « champion de judo » surgissait ainsi sur les écrans d'ORT et de RTR, les chaînes de télévision publiques. Ces images le montraient en kimono, le visage et le corps tendus dans l'effort, agile et rapide comme l'éclair, envoyant au* ***tapis*** *tous ces adversaires. Quelques semaines auparavant, le premier ministre s'était embarqué à bord d'un avion Mig pour faire des looping. POURTANT les Russes n'en savent guère plus sur leur nouveau président par intérim. »* (extrait tiré de l'<article id="LM10-d682945p4" rub="HOR")

Dans l'article d'où est extrait cette occurrence, il y a un peu plus de 3% de mots du sport.

Au vu du contexte général nous pourrions en déduire qu'il ne s'agit pas d'un emploi sportif. Pourtant, le mot *judo* permet de savoir que, même s'il s'agit d'une image, l'emploi est bel et bien sportif. Une analyse du contexte permet de le savoir mais le calcul de pourcentage n'est pas adapté pour détecter de tels cas. Nous pouvons donc dire que le contexte permet de détecter un emploi spécialisé, mais que des recherches plus poussées doivent être effectuées sur les différents types d'emplois pour ajuster le calcul qui permet à l'ordinateur de trancher.

# CHAPITRE V : PERSPECTIVES

## I. AMÉLIORATION DE LA DÉTECTION DE CONTEXTE SPÉCIALISÉ

Dans cette partie, nous revenons sur les différents éléments mis en place pour vérifier l'hypothèse de travail et envisageons des solutions pour avoir de meilleurs résultats.

### 1. Amélioration du lexique (LexSpo)

Le lexique mis en place pour la vérification de l'hypothèse permet de détecter jusqu'à 92% d'occurrences spécialisées. Pour améliorer ces scores, plusieurs options doivent être explorées.

#### a. Ajouter des noms propres

Dans un premier temps, ainsi que nous l'avons mentionné dans la partie II. 2. du chapitre III (cf. p. 36), une réflexion sur les noms propres et leur apport pour la détection des contextes spécialisés pourrait être menée. Les noms propres font partie du vocabulaire d'une langue de spécialité. En théorie, ils participent à la création du contexte. Dans certains cas ils sont même les seuls éléments du contexte qui permettent d'identifier un emploi sportif. Par exemple, voici un article tiré du corpus :

<article id="LM10-d683491p2" rub="SPO">

LE **FOOTBALLEUR** ARGENTIN DIEGO MARADONA, 39 ans, a été hospitalisé mardi 4 janvier dans une clinique privée de la station balnéaire de Punta del Este. Selon l'équipe médicale de l'établissement, l'ancien international, sacré récemment sportif du siècle en Argentine, souffre d'hypertension et d'arythmie. « Son état est stable et il est un peu excité », a déclaré à l'AFP par téléphone un médecin de la clinique Cantegril, ajoutant que l'épouse du footballeur, Claudia Villafane, se trouvait avec lui.L'agent de Diego Maradona, Guillermo Coppola, a affirmé que l'hospitalisation de l'Argentin n'est pas liée à des problèmes de drogue. « Diego est arrivé à la clinique en conduisant sa voiture. Il n'y a pas eu besoin d'ambulance, il a même fait des photos dans la journée », a-t-il déclaré à la radio argentine Mitre.

Le terme *footballeur* (en gras) désigne bien un joueur de football. Mais avec la méthode envisagée, rien ne permet de savoir qu'il s'agit d'un emploi sportif à part le nom du joueur, *Diego Maradona*. Ajouter les noms des sportifs, des clubs, des équipes, des noms de bateaux, de chevaux,... peut permettre de faire monter le pourcentage de certaines occurrences qui apparaissent dans un contexte un peu pauvre en mots spécialisés.

### b. Tenir compte de la nature des mots du lexique

Certains mots du sport, comme *pas* et *par* que nous évoquions plus haut (cf. p. 36) sont des noms, mais ils ont également des homonymes qui apparaissent beaucoup plus fréquemment et qui sont, eux, des adverbes et des prépositions. C'est pourquoi nous avons décidé de ne pas les inclure dans LexSpo, pour ne pas fausser les calculs. Nous recherchons des occurrences en contexte dans un corpus au format étiqueté, c'est-à-dire que nous avons accès à la classe grammaticale de chaque mot du corpus. Nous envisageons d'étiqueter les mots de LexSpo pour pouvoir y ajouter des mots ambigus comme les deux mentionnés tout en limitant le risque que cela ne fausse les calculs de pourcentage. Nous préciserions en effet au programme que les lemmes doivent correspondre, mais que les classes grammaticales doivent également correspondre.

Une bonne manière de tester l'efficacité de LexSpo serait de lancer le programme sur un autre corpus, afin de s'assurer que les résultats ne sont pas uniquement dus au fait que le lexique est majoritairement constitué de mots du sous-corpus LMsport.

## 2. Travail sur la taille de la fenêtre contextuelle

Dans ce travail la fenêtre contextuelle correspond à l'ensemble d'un article. Si les articles sont longs, ou si le corpus utilisé n'est pas journalistique, la taille de la fenêtre peut devenir un problème. Nous disions en effet dans le chapitre III (cf. partie I. 3. p.32), que les articles journalistiques ont souvent un sujet principal qui court sur tout l'article, ce qui permet de larges fenêtres contextuelles. Mais l'exemple présenté en conclusion des manipulations (cf. p.57) nous montre qu'il arrive qu'un emploi soit isolé dans un paragraphe ou une phrase, et analyser l'article entier devient contre-productif. C'est pourquoi il serait pertinent de réfléchir à la taille de fenêtre contextuelle la plus adéquate. Une façon de procéder serait de reprendre la procédure et les manipulations effectuées dans le chapitre IV pour établir le seuil optimal de la proportion de mots de LexSpo dans le contexte, mais cette fois en fixant à chaque fois une taille de fenêtre contextuelle différente, pour voir laquelle permet d'obtenir les meilleurs résultats. Nous avions envisagé en premier lieu de faire cette manipulation après avoir fixé le seuil de mots spécialisés, mais ces deux paramètres peuvent ne pas être indépendants et nous ne pouvons pas toucher à la fenêtre contextuelle sans modifier les pourcentages, donc potentiellement la valeur seuil. C'est pourquoi il faudrait reprendre la procédure du début à chaque fois.

Les paragraphes développés jusqu'ici dans ce chapitre visent à proposer des pistes de réflexion pour mener à son terme le processus de vérification de l'hypothèse, et s'assurer qu'analyser le contexte est bien suffisant pour détecter une déterminologisation. Les manipulations effectuées dans ce travail de recherche tendent en effet à prouver qu'il y a un lien

entre contexte et emploi et qu'il est possible de déduire automatiquement un emploi spécialisé en analysant son contexte, mais il faut maintenant s'assurer que les déterminologisations sont détectables, elles aussi. Le cas échéant, nous détaillons dans la partie suivante quelques pistes pour la poursuite de la méthode envisagée.

# II. POURSUITE DE LA MÉTHODE

Dans cette partie nous expliquons les différentes étapes que nous envisageons pour mettre au point un programme de détection automatique de cas de déterminologisation.

## 1. Créer un lexique spécialisé.

Il est possible de s'inspirer de LexSpo pour créer le lexique spécialisé du sport, mais l'objectif est que sa version achevée puisse s'adapter à différents types de corpus. Nous avons envisagé de récupérer les mots des lexiques sport du Wiktionnaire[9] ou les titres de pages du portail sport de Wikipédia[10] par exemple. Ces ressources cumulent plusieurs avantages : elles sont libres de droit et mises à jour régulièrement. Cette dernière caractéristique laisse penser que le lexique restera représentatif de la langue de spécialité. En effet, si cette méthode est mise en place sur la durée, l'un des enjeux du lexique est qu'il doit rester à jour afin que le contexte soit détecté le plus efficacement possible. Ceci est d'autant plus vrai si l'utilisation de noms propres se révèle pertinente et s'il contient des noms de personnalités sportives, car de nouveaux sportifs se font connaître très régulièrement. Et si le programme analyse des articles de presse, les noms évoqués seront vraisemblablement des noms qui font l'actualité. Si le lexique n'est pas à jour, la détection sera moins efficace. Il est également nécessaire d'automatiser la gestion du contenu du lexique, pour des questions d'efficacité.

## 2. Élaboration d'un score de spécialisation.

Pour vérifier la validité de l'hypothèse, nous nous sommes appuyé uniquement sur un calcul de pourcentage. Il serait pertinent de réfléchir à un score un peu plus élaboré. Une étude pourrait être menée sur l'importance de certains types de mots par rapport à d'autres dans l'élaboration d'un contexte spécialisé pour mettre en place une pondération qui accorde plus de poids à ces mots là. Par exemple, si nous conservons une fenêtre contextuelle large, nous pourrions donner plus d'importance aux mots spécialisés qui se trouvent proches du mot-cible analysé. L'objectif est ici de limiter les erreurs dues à l'apparition d'emplois spécialisés dans des contextes pauvres en vocabulaire spécialisé, et vice versa.

9 https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Sport

10 https://fr.wikipedia.org/wiki/Portail:Sport

## 3. Que faire avec les emplois non sportifs ?

Une fois les emplois non sportifs détectés, il ne s'agit pas de tous les signaler au lexicographe au fur et à mesure. En effet, un dictionnaire n'est pas mis à jour lorsqu'une occurrence apparaît dans un contexte inhabituel. Il faut que le schéma se répète et prenne une certaine place dans la langue pour être recensé. Des études sur les mots ayant des emplois déterminologisés pourraient être menés, afin de déterminer la proportion d'emplois spécialisés par rapport à la proportion d'emplois despécialisés pour un mot donné. L'objectif de ce travail est de savoir s'il est possible de fixer un pourcentage ou un nombre seuil d'emplois inhabituels pour chaque mot analysé, au delà duquel le lexicographe est alerté et peut analyser les fichiers de sortie afin de déterminer si les occurrences extraites correspondent bien à des emplois déterminologisés du mot. Nous pouvons aussi imaginer que le lexicographe détermine lui même à partir de combien d'occurrences inhabituelles d'un mot il souhaite être averti pour vérification.

Les pistes développées ici sont destinées à, d'une part, s'assurer que l'hypothèse de départ est assez solide pour développer la méthode envisagée, et d'autre part à identifier les aspects les plus importants à mettre en œuvre pour espérer voir la méthode fonctionner.

# CONCLUSION

Au cours de ce mémoire, nous nous sommes penché sur le phénomène de la déterminologisation, qui correspond au glissement d'un emploi de spécialité vers un emploi plus général. Nous avons montré en quoi être capable de repérer ce phénomène automatiquement pouvait être profitable à la lexicographie assistée par ordinateur. Nous avons choisi de nous intéresser au domaine sportif, pour lequel nous supposions des prédispositions à la déterminologisation du fait de son statut populaire et accessible. Nous pensions en effet que ces caractéristiques donnaient lieu à une communication importante, susceptible d'entraîner une réappropriation du vocabulaire par les locuteurs. Nous tenons à nuancer ici ce présupposé. En effet, le caractère populaire du sport fait que le vocabulaire utilisé dans les corpus journalistiques est déjà très courant. Les journaux n'étant pas destinés uniquement à un public professionnel, les rédacteurs doivent s'efforcer de rester compréhensibles si bien que le vocabulaire y est peu technique. Bien sûr il relève du domaine du sport, mais il est difficile de savoir dans quelle mesure il est, ou non, déterminologisé puisqu'il est souvent considéré comme courant et maîtrisé par une majorité de locuteurs. Un emploi despécialisé issu de la médecine ou de l'informatique aurait peut-être été plus facile à repérer puisque cela reste des domaines de spécialité dont la langue est maîtrisée par les spécialistes. Peut-être aurions nous également eu moins de difficultés en nous concentrant sur un sport en particulier et sur tout son vocabulaire technique plutôt que de rester sur le domaine sportif en général.

Nous avons tout de même isolé quelques exemples, qui nous ont permis de poursuivre et de mettre au point une méthode de détection automatique de la despécialisation fondée sur l'hypothèse que le contexte permet de déterminer le type d'emploi d'un mot. Il s'est agi par la suite de vérifier la validité de cette hypothèse à l'aide du dictionnaire GLAWI et d'un corpus d'articles *du Monde*. Nous avons observé plusieurs échantillons de mots ayant des emplois sportifs et analysé automatiquement leur contexte en comparant les mots de ces contextes à un lexique spécialisé. La méthode mise en place pour vérifier l'hypothèse ainsi que les résultats obtenus à partir des différentes manipulations montrent qu'il existe bien un lien entre l'emploi d'un mot et le type de contexte dans lequel il apparaît. Certains ajustements doivent être effectués pour pouvoir conclure avec certitude, mais il semble qu'il soit possible de détecter un type d'emploi à partir du contexte dans lequel il apparaît. Cependant, la méthode de calcul utilisée doit être améliorée pour prendre en compte les cas particuliers où une occurrence sportive apparaît dans un contexte pauvre en mots spécialisés.

Si l'hypothèse était vérifiée et que le reste de la méthode était mise en place avec succès, elle pourrait être implémentée sur un dictionnaire pour aider les lexicographes à mettre à jour les articles de dictionnaire.

# RÉFÉRENCES

## Références bibliographiques


Atkins, B.T.S et Rundell M. (2008). *The Oxford Guide to Practical Lexicography*. New-York :Presses de l'université d'Oxford.

De Bessé, Bruno. (2000). Le domaine. Dans Béjoint, H et Thoiron P. (dir.). *Le sens en terminologie*, 198–217. Lyon: Presses Universitaires de Lyon.

Cabré, M. T. (1998). *La terminologie : théorie, méthode et applications*. Ottawa : Presses de l'Université d'Ottawa et Armand Colin.

Galisson, R. (1978). *Recherches de lexicologie descriptive : la banalisation lexicale. Le Vocabulaire du football dans la presse sportive. Contribution aux recherches sur les langues techniques*. Paris : Nathan.

Guilbert L. (1975). *La créativité lexicale*. Paris : Larousse.

Harris, Z.S. (1968). *Structures mathématiques du langage*. Paris : Dunod.

Humbert-Droz, J. (2004). *Le passage de termes d'une langue de spécialité à la langue générale : le cas du domaine spatial* (Mémoire de master, Université de Genève). Repéré à https://archive-ouverte.unige.ch/unige:41152

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133. Repéré à http://www.kilgarriff.co.uk/Publications/2001-K-CompCorpIJCL.pdf

Kocourek, R. (1991). *La langue française de la technique et de la science : vers une linguistique de la langue savante*. Wiesbaden : O. Brandstetter Verlag.

Meyer, I. et Mackintosh, K. (2000). "L'étirement" du sens terminologique : aperçu du phénomène de la déterminologisation. Dans Béjoint, H et Thoiron P. (dir.). *Le sens en terminologie*, 198–217. Lyon: Presses Universitaires de Lyon.

Pruvost, J. (2000). *Dictionnaires et nouvelles technologies*. Paris : PUF.

Rondeau, G. (1981). *Introduction à la terminologie*. Montréal: Gaëtan Morin.

Sajous, F., et Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. Dans *Proceedings of the eLex 2015 conference.* 405-256. Repéré à https://halshs.archives-ouvertes.fr/halshs-01191012/.

Sinclair, J. (2004). *Trust the Text, language, corpus and discourse*. Londres :R Taylor and Francis e-library.

——— (2005). Corpus and Text – Basic Principles. Dans Wynne M. (dir.). D*eveloping Linguistic Corpora: a Guide to Good Practice*, 1-16. Oxford : Oxford Books.

## Références sitographiques

ATILF. (2002). TLFi . Repéré à http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?11;s=1169864775;r=1;nat=;sol=0;

Sajous, F. (2015). GLAWI : GLAFF et WiktionaryX. Repéré à http://redac.univ-tlse2.fr/lexiques/glawi.html

## Déclaration sur l'honneur de non-plagiat

(à joindre au mémoire à la fin du document)

Je soussigné.e,

Nom, Prénom : DELVENNE Léa

Régulièrement inscrit.e à l'Université de Toulouse II Jean Jaurès

N° étudiant : 21604362

Année universitaire : 2016-2017

certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à : Toulouse

Le : 17/06/2017

Signature :