

Enrichir le balisage de corpus footballistiques pour en augmenter le pouvoir documentaire¹

Nathalie Gasiglia

Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille

nathalie.gasiglia@univ-lille3.fr

Résumé. La présente contribution ambitionne de revenir sur l'élaboration, durant la dernière décennie, de deux corpus de commentaires footballistiques – l'un d'oral transcrit et l'autre de sources écrites – de taille modeste mais constitués afin qu'ils présentent un haut rendement exploratoire. L'objectif de ce retour est de réexaminer les choix de balisage XML mis en œuvre au sein de chacun et d'étudier la pertinence d'une annotation plus fine de certains phénomènes non encore traités. Après avoir présenté le contexte qui a motivé l'élaboration de ces corpus puis leurs contenus, nous exposons les éléments majeurs du balisage de chacun, en valorisant ce qui les différencie, avant d'entrer plus finement dans leurs données et d'analyser l'impact des modalités ou des moments de production des énoncés sur leur forme, et conséquemment la pertinence d'une annotation de ces paramètres dans le balisage. Si les deux corpus ont à l'heure actuelle un haut rendement exploratoire du fait de leur thématisation et de la sélection de commentaires, donc de productions d'un ensemble de locuteurs spécialisés qui s'adressent à un large public, la spécificité des situations d'énonciation propres aux énoncés oraux ou écrits de chaque corpus a un impact sur la nature des données observables au sein de chacun. Ainsi, d'une certaine manière, selon le média de diffusion, les commentaires qui permettent au public de suivre les matchs qu'il ne voit pas ne l'informent pas de manière équivalente. Nous apprécierons comment le balisage XML peut faciliter les analyses des discours et les études lexicales au sein de ces corpus.

Abstract. This paper revisits the ten-year-long elaboration of two football commentary corpora, one transcribed from spoken language, the other based on written sources. The two corpora are of small size but with high exploratory efficiency. The goals of this new study are to reexamine the used XML markup choices of each corpus and to investigate the relevance of a more detailed tagging for certain phenomena not yet treated. After presenting the context motivating the elaboration of these corpora then their contents, the major markup elements of each are displayed, highlighting their differentiating features, then data is explored in more detail and the impact of utterances production modalities or moments on their form, and finally the relevance of such parameters' tagging in the markup, are analyzed. Both corpora have a high exploratory efficiency due to their thematic character and commentaries selection, that is utterances of a certain number of specialized speakers addressing a large audience. However, the specificity of utterances production situations has an impact on the nature of observable data in each corpus: depending on the medium, the commentaries, allowing the audience to follow the matches without seeing them, do not inform in equivalent manner. I will estimate how XML markup may facilitate discourse analysis and lexical studies in these corpora.

Je vais revenir sur l'élaboration, depuis le début des années 2000, de deux corpus thématiques – l'un d'oral transcrit et l'autre de sources écrites – de taille modeste mais constitués afin qu'ils offrent un haut rendement exploratoire.

L'objectif de ce retour est, d'une part, de repérer quelques similitudes et différences observables entre ces corpus – dont en particulier celles liées aux modalités ou aux moments de production des énoncés de chacun –, et, d'autre part, d'étudier la pertinence d'une annotation plus fine de certains objets.

Après avoir présenté le contexte qui a motivé l'élaboration de ces corpus puis leurs contenus, j'évoquerai donc les éléments majeurs de la structuration XML de chacun, puis l'introduction d'annotations complémentaires.

1 Contexte de création des deux corpus thématiques à haut rendement considérés

Au début des années 2000, afin de proposer aux lexicographes français de pallier leur manque de ressources accessibles, Pierre Corbin (UMR STL, Univ. Lille) et moi avons introduit la notion de "corpus thématique à haut rendement", auprès d'étudiants en formation dans ce qui est aujourd'hui le master Lexicographie, Terminographie et Traitement Automatique des Corpus (LTTAC)².

Le contexte était alors le suivant : dans les maisons d'édition françaises, les lexicographes n'avaient pas eu les moyens ou le souhait de suivre leurs homologues britanniques. Ceux-ci avaient été les premiers, deux décennies plus tôt, à promouvoir la notion de corpus "de référence"³ et à valoriser les exemples issus de corpus⁴, et ils s'engageaient dans l'exploitation de corpus Web, en particulier avec Adam Kilgarriff et son Sketch Engine⁵.

À Lille, notre projet était de montrer que, pour les éditeurs français, une alternative au corpus de référence – dont la France ne semblait pas devoir se doter –, ou aux corpus Web – à l'égard desquels une partie des chefs de projets avaient une réticence mal surmontable –, pouvait être l'exploitation de "corpus thématiques à haut rendement", c'est-à-dire des corpus de taille modeste mais constitués par une sélection stricte de documents primaires et informant sur un ensemble circonscrit d'usages.⁶

Afin d'illustrer l'intérêt de ce type de corpus, nous avons choisi de nous intéresser aux commentaires de matchs de football comme premier terrain d'expérimentation, du fait d'une part de la richesse lexicale de ces énoncés, et d'autre part de leur centrage sur des descriptions d'actions de jeu qui sont assez étroitement liées aux règles de ce sport.

L'exploitation de ce type de narrations me semblait transposable à d'autres espaces expérientiels – en fait beaucoup de domaines techniques dont les actions sont elles aussi codifiées par des protocoles opératoires ou d'expérimentation –, mais les productions langagières de certains d'entre eux peuvent ne pas bénéficier de la même diffusion et donc être moins accessibles aux analystes externes.

Un premier corpus de commentaires footballistiques radiophoniques transcrits de type multiplex a été élaboré à partir de 2002-2003 avec nos étudiants. Il réunit des commentaires de matchs de football radiodiffusés dans le cadre de multiplex, principalement ceux d'*Europe 1*, qui ont été transcrits et dont les transcriptions ont été structurées en XML⁷. Le balisage permet de délimiter les interventions dans les tours de parole en identifiant les locuteurs et annoter une sélection de particularités linguistiques (comme des prononciations remarquables) et d'entités nommées (noms de joueurs, de clubs, de stades, etc.)⁸. Le corpus

compte 200 000 mots-occurrences environ, ce qui correspond à 10 soirées de multiplex découpées en près de 6 000 interventions au total (de 400 à 850 par soirée).

Le choix des multiplex est lié à la simultanéité des commentaires dans différents stades, qui garantit une bonne densité de descriptions d'actions de jeu (même si certaines d'entre elles sont plus valorisées que dans les commentaires continus).

Plus récemment, des comptes rendus de matchs en direct sont apparus sur Internet et ont alimenté un second corpus, lui aussi structuré en XML, qui a été constitué en 2011-2012 avec Hans Paulussen⁹ (ILT, K.U.Leuven, site de Courtrai). Il s'agit d'un corpus de taille modeste pour de l'écrit (150 commentaires de matchs, soit 270 000 mots-occurrences environ) élaboré afin de tester des procédures de constitution et d'annotation semi-automatiques : téléchargement des sources HTML sur le site *lequipe.fr*, conversion en XML en conservant maximale la richesse informationnelle originale, puis lemmatisation et annotations (morphosyntaxique et en chunks) des textes des interventions avec la chaîne de traitement Macaon¹⁰.

Le choix de réunir en corpus des commentaires footballistiques s'articulait à deux souhaits :

(i) celui de promouvoir la notion de "lexique spécialisé de masse", un lexique utile pour une expression d'une relative technicité, mais destiné à un très large public du fait des modes de diffusion des énoncés dans lesquels il figure : émissions radio- ou télédiffusées¹¹, ou écrits de presse imprimée ou en ligne ;

et (ii) celui, dans le cadre de la formation de lexicographes, de fonder les descriptions d'unités lexicales sur des explorations de corpus.

Ce dernier souhait était induit par l'observation de la faible prise en compte, dans les dictionnaires généraux et spécialisés, de ces emplois spécialisés de large diffusion (ce qu'ont montré les publications listées en n. 6).

2 Présentation des corpus footballistiques créés

Les deux corpus footballistiques créés se caractérisent par la haute fréquence, d'une part, des descriptions de gestes techniques et, d'autre part, des entités nommées et des expressions de localisation, qui s'avèrent être de bons points d'accès aux expressions qui décrivent les actions de jeu :

1) Extraits d'interventions du Corpus footballistique de multiplex : désignations de joueurs en PETITES CAPITALES, d'équipes en PETITES CAPITALES SOULIGNÉES et de lieux en **gras**

• « [...] à l'instant là PÉDRON pour euh BAKARI et ULRICH RAMÉ est obligé de sortir de ses **dix-huit mètres** et de dégager le ballon de la tête d'une tête plongeante c'est dire si l'on voit des occasions côté bordelais on a eu l'ébauche d'une esquisse d'une bonne combinaison entre PAULETA et DARCHEVILLE qui s'est terminée par un tir du gauche de SAVIO WARMUZ impeccable dans **les buts lensois** a dévié ce ballon **en corner** et en face c'était une euh un une deux UTAKA PÉDRON qui a abouti à une certaine menace vis-à-vis de ULRICH euh RAMÉ c'était deux bonnes occasions de but dans un match qui se déroule parfaitement [...] »

• « [...] il y a quelques instants pour LES GIRONDINS une balle parfaitement distillée longue en profondeur **dans l'axe lensois** lancée par euh CHRISTOPHE DUGARRY à l'intention et à l'attention de PAULETA ce diable de PAULETA qui met toujours le nez et le pied à la fenêtre a failli récupérer ce ballon et ça a été un sauvetage extraordinaire acrobatique du DÉFENSEUR LENSOIS BAK qui a enlevé le ballon alors que l'on le voyait déjà **au fond de la cage de GUILLAUME WARMUZ** [...] »

2) Extraits d'interventions du Corpus footballistique Web (*lequipe.fr*) au fil du match : noms de joueurs en PETITES CAPITALES et désignations de lieux en **gras**

• « [coup franc] Excentré **sur la gauche**, DZAGOEV propose une frappe tendue **dans la surface**. BÉRIA écarte de la tête et concède le premier corner. »

• « LILLE réagit **sur le côté gauche** avec MOUSSA SOW, qui obtient le premier corner lillois. »

L'exploration outillée de ces corpus pour nourrir des analyses lexicales valorise par ailleurs bien les spécificités d'emplois particuliers d'items d'usages plus larges.



Comme nous l'avons succinctement indiqué ailleurs (cf. Corbin (2008) et Gasiglia (2010) notamment) à la suite de Deulofeu (2000), ces corpus fournissent également un bon matériau pour des analyses de discours, puisqu'ils regroupent :

– pour celui des multiplex : des énoncés transcrits qui sont spontanément formulés, et qui sont produits soit en direct, au moment où les actions décrites se déroulent, soit en léger différé, au moment où l'animateur qui est en studio interroge chacun des reporters présents dans les stades,




– et pour celui de *lequipe.fr* : des textes brefs rédigés en léger différé par un rédacteur unique qui peut organiser mentalement son exposé avant de le saisir.

Il convient toutefois d'observer que certaines rédactions en ligne se font en deux temps : dans l'exemple suivant, à 17h43, l'annonce du but a brièvement été faite presque en direct, puis, à 18h12, elle a été reformulée et l'indication de la minute de jeu à laquelle le but a été marqué a été corrigée. Le corpus étant constitué des textes disponibles après les matchs, seules les productions réécrites y sont attestées. L'illustration proposée ci-dessous pour ce phénomène n'est donc pas un texte du corpus, mais un autre, observé et relevé plus récemment. Il est relatif au match Lille / Valenciennes du 27 octobre 2012, lu en ligne à 17 h 43 et 18 h 12 (cf. <http://www.lequipe.fr/Football/match/267227>).

1) 17 h 43 :

| | | |
|----|---|---|
| 38 |  | But de Payet ! |
| 32 | | Les visiteurs perdent presque tous les duels, surtout dans l'entrejeu. Lille montre pour le moment beaucoup d'agressivité et de volonté, à l'image de Rozehnal, qui ne joue pourtant pas à son poste. |
| 36 |  | A droite, Payet prend le coup de pied de coin. Au second poteau, Rozehnal ose la reprise en force. Au-dessus. |

2) 18 h 12 :

| | | |
|----|---|---|
| 43 |  | Le centre de Danic est mal négocié par la défense du LOSC. Mais Gil n'en profite pas. |
| 42 | | Ducourtieux essaie de montrer l'exemple avec un raid sur le côté droit. Le capitaine obtient un corner. |
| 41 |  | Lille assomme Valenciennes avant la pause au terme d'un joli mouvement collectif et d'une chevauchée plein axe de Kalou, qui, au bout de sa course, a décalé vers Payet. Ce dernier trompe Penneteau avec un tir croisé du droit à ras de terre. |
| 36 |  | A droite, Payet prend le coup de pied de coin. Au second poteau, Rozehnal ose la reprise en force. Au-dessus. |

L'observation de ces réécritures ponctuelles contraint à poser que le corpus Web constitué ne retient que le dernier état des commentaires et non strictement ce que les internautes ont pu lire au fil des matchs. La prise en compte de ces reformulations (sous la forme de segments alignés) pourrait être intéressante pour l'analyse des discours journalistiques, mais elle compliquerait considérablement le mode de constitution du corpus puisqu'il faudrait surveiller ces réécritures qui ne laissent aucune trace *a posteriori*.

3 Éléments majeurs du balisage des deux corpus considérés et comparaison des contenus de ceux-ci

Venons-en maintenant aux éléments majeurs du balisage des deux corpus footballistiques. S'ils sont structurés en XML, la nature des commentaires et leur mode de collecte induisent des différences de balisage.

Dans le corpus de multiplex, les énoncés transcrits sont balisés en prises de paroles pour lesquelles les énonciateurs sont typés et identifiés. Ces interventions s'enchaînent ou, parfois, se superposent au sein des tours de paroles, ce qui motive le repérage de leurs chevauchements. Dans chaque intervention, comme dans l'extrait ci-dessous, un certain nombre d'objets sont balisés : des entités nommées de diverses natures (joueurs, clubs, villes, stades, etc.) et des prononciations inattendues (remarquables ou accidentelles).

```

<!-- [réduction] -->
du score ici au stade
<ENTITY TYPE-OF-ENTITY="stade">Jean Laville</ENTITY>
<ACCIDENTAL-PRONUNCIATION>
  <TRANSCRIPTION-OF-ACCIDENTAL-PRONUNCIATION>
    por
  </TRANSCRIPTION-OF-ACCIDENTAL-PRONUNCIATION>
  <SPELLING-OF-ACCIDENTAL-PRONUNCIATION>
    pour
  </SPELLING-OF-ACCIDENTAL-PRONUNCIATION>
</ACCIDENTAL-PRONUNCIATION>
<ENTITY TYPE-OF-ENTITY="equipe">
  Créteil
</ENTITY>
c'est
<!-- [...] -->
oui ballon bordelais avec
<REMARKABLE-UTTERANCE>
  <TRANSCRIPTION-OF-REMARKABLE-UTTERANCE>
    affolo
  </TRANSCRIPTION-OF-REMARKABLE-UTTERANCE>
  <SPELLING-OF-REMARKABLE-UTTERANCE>
    affolo
  </SPELLING-OF-REMARKABLE-UTTERANCE>
  <ANALYSIS-OF-REMARKABLE-UTTERANCE>
    apocope supposée de "affolement"
  </ANALYSIS-OF-REMARKABLE-UTTERANCE>
</REMARKABLE-UTTERANCE>
de la défense troyenne
<!-- [...] -->

```

Le balisage de ce corpus avait été conçu en nous inspirant des recommandations de la TEI, mais sans nous y conformer. Comme ce corpus est manipulé au moyen de différentes transformations XSLT écrites pour lui, la standardisation de son balisage, qui impliquerait la révision de l'ensemble des outils développés, serait coûteuse, mais elle est envisagée dans le cadre d'un enrichissement à venir des annotations.

Pour le corpus Web, l'exploitation de comptes rendus de matchs en ligne a permis, lors de la conversion du balisage HTML en XML, de tirer profit d'indications déjà codifiées pour chaque intervention : le temps de jeu écoulé depuis le début d'un match et le nom des pictogrammes associés à certains événements du jeu.

Le produit de cette conversion pouvait déjà être exploré, mais, comme dans le corpus radiophonique, le balisage des entités nommées y a en outre été entrepris¹², comme dans l'extrait suivant¹³ :

```

<div>
  <p type="minn">78</p>
  <p type="coup_franc"> </p>
  <p>
    <persName role="milieu" key="OGCN">Hellebuyck</persName>
    frappe à ras de terre à vingt-cinq mètres.
    <persName role="gardien" key="FCGB">Carrasso</persName>
    capte sur sa ligne.
  </p>
</div>

```

Le texte des interventions a par ailleurs été annoté avec la chaîne de traitement Macaon, développée au LIF (à Marseille, cf. n. 10), mais la qualité des lemmatisations, des annotations morphosyntaxiques et des découpages en chunks obtenus ne peut pas être analysée dans l'espace dévolu à cette contribution.

3.1 Comparaison des corpus : approche globale

Pour débiter la comparaison des deux corpus, observons que, dans les multiplex, ce sont plusieurs matchs qui sont commentés en simultané, avec une valorisation des éléments les plus saillants de ce qui se déroule sur les différents stades, alors qu'en ligne un match unique est commenté en continu.

Par ailleurs, les reporters des multiplex assistent aux matchs dans les stades, mais probablement pas les commentateurs de *lequipe.fr*, puisque certains commentent plusieurs matchs (dans des stades éventuellement éloignés) dans une même journée. En conséquence, dans le cadre d'une étude intertextuelle, il serait intéressant de déterminer comment la source d'information des seconds a orienté la sélection des séquences de jeu décrites et la formulation des interventions. Cette source est probablement télévisuelle, puisqu'il arrive qu'un ralenti soit évoqué dans le commentaire rédigé, comme c'est le cas dans l'extrait ci-dessous :

« D'après le ralenti, le buteur était en position de hors-jeu au départ de l'action... »

À partir des corpus dont nous disposons, nous pouvons globalement comparer les commentaires des différents médias, ou le faire pour un match, en sélectionnant les interventions d'un reporter d'un multiplex et celles d'un match en ligne. Toutefois les corpus contenant des commentaires des saisons 2002-2003 pour l'un et 2011-2012 pour l'autre, il ne nous est pas possible de savoir pour un même match si les mêmes événements sont décrits simultanément et dans les mêmes termes à la radio et en ligne.

Les tailles relatives des commentaires des deux corpus, elles, sont comparables, avec, pour un match, 1 700 à 1 800 mots en moyenne dans chaque cas, mais un reporter peut plus ou moins intervenir lors d'une soirée de multiplex, les rapports entre les productions des différents énonciateurs pouvant être de 1 à 5, alors qu'en ligne les variations sont plus réduites (de 1 à 1,5).

3.2 Comparaison des corpus : contenu des interventions

Concentrons-nous maintenant sur les interventions elles-mêmes. Dans les corpus, celles qui sont internes aux tours de parole et celles en ligne partagent la propriété d'être des éléments fondamentaux pour la délimitation de ce qui est exprimé par un locuteur à un moment donné, mais elles diffèrent par ailleurs. Les contextes discursifs semblent avoir une réelle influence sur les modalités d'expression concernant, notamment, le résultat d'une action de jeu ou les indications temporelles ou de localisation.

3.2.1 Résultat d'une action de jeu

À la radio, les actions aboutissant à un but sont souvent narrées deux fois dans une même intervention : quand le reporter prend la parole à l'occasion d'un but, il introduit son propos par la description de l'action qui a amené celui-ci, puis il revient sur la séquence de jeu qui a conduit à elle, avant de la répéter en la reformulant éventuellement :

- « ouverture du score au stade Abbé-Deschamps **c'est Kapo de la tête qui vient de marquer le but** à la suite d'un coup franc de Lachuer donné dans la surface de réparation Kapo a sauté plus haut que les défenseurs de Sedan et **de la tête il a mis le ballon hors de portée de Patrick Regnault** [...] »
- « [...] l'ouverture du score pour Lille **but inscrit par Hector Tapia** oh c'est parti de très loin un centre venu de la droite par euh Tafforeau qui est parti là-bas au deuxième poteau une reprise de Philippe Brunel qui était tout juste sur la ligne euh de but une remise en retrait pour euh **Hector Tapia à l'affût dans la surface de réparation pour reprendre ce ballon de la tête une tête piquée directement dans la cage de Ronan Le Crom** [...] »

Ceci ne s'observe pas à l'écrit, où il n'y a pas de répétition mais plutôt des implicites textuels. En effet, sur le Web, l'introduction d'une intervention, par exemple, par le pictogramme du ballon (qui symbolise un but) fait qu'il n'est pas utile de dire explicitement ce qu'il code et que le commentateur peut se concentrer sur l'action qui a amené le but :

« Le Gym se détache au bout d'une action limpide. Sur un dégagement rapide d'Ospina, Guié Guié temporise et décale vers Meriem. Celui-ci trouve Hellebuyck, qui perfore la défense bordelaise et finit habilement du pied gauche. »

Mais l'implicite n'est pas non plus une règle absolue :

« Profitant d'un mauvais remplacement de Savic, Mouloungui, servi entre deux adversaires, accélère dans la surface girondine et **ouvre la marque** avec un tir du pied gauche sans angle. L'attaquant s'est peut-être aidé de la main... »

3.2.2 Indications temporelles

À la radio, les indications temporelles sont omniprésentes, peut-être parce que, les matchs ne débutant pas tous strictement en même temps, chaque reporter doit indiquer soit le temps de jeu écoulé au moment où il parle, s'il relate un événement présent, soit celui où une phase de jeu s'est produite, s'il la restitue en différé. Il peut par ailleurs situer ce qui s'est déroulé et qu'il raconte soit par rapport à son énonciation en employant des expressions comme « à l'instant » ou « il y a quelques minutes », soit, occasionnellement, par rapport à l'intervention du précédent commentateur.

Indications de temps de jeu écoulé (**en gras**) et de repérages temporels approximatifs, situant ce qui est raconté par rapport au moment de l'énonciation (*en italique gras*) ou par rapport à l'intervention du commentateur précédent (*en italique maigre*) :

• « [...] deux occasions très franches pour les joueurs de Thierry Goudet **à la cinquième minute** Celdran a alerté euh Cousin Cousin qui s'est débarrassé de son garde du corps mais qui s'est heurté à Marichez bien sorti à sa rencontre sur le corner qui a suivi Léon Bollée a cru au but mais la tête de Pancrate a été repoussée sur la ligne une timide occasion également à l'actif des Chamois niortais c'était d'ailleurs **dès la première minute** un coup franc de vingt-cinq mètres Darbelet qui a décalé euh Foulon dont la frappe a été repoussée à hauteur du point de penalty par la défense mancelle [...] »

• « et **après six minutes** toujours euh zéro à zéro mais nette domination des joueurs de Créteil qui se sont procuré deux occasions très franches notamment par l'intermédiaire de ce jeune russe Kossonogov prêté par les Girondins de Bordeaux à Créteil et qui a **tout à l'heure** devancé la sortie de Tingry à la limite de la surface de réparation et d'une pichenette eh bien son ballon est passé tout près des buts qui étaient donc vides et **quatre minutes plus tard** ce même Kossonogov eh bien a donné un ballon en or à Sébastien Dallet le capitaine de Créteil qui se présente tout seul face à Tingry et cette fois euh Tingry sauve son camp en expédiant ce ballon en corner côté rémois une seule petite occasion un tir en pivot d'Olivier Pickeu de vingt mètres qui est passé quand même à côté des buts de Legrand donc **pour l'instant** zéro zéro »

• « **à l'instant** c'est Grégory Lacombe qui a tenté sa chance ah c'était loin quand même très loin pour euh le petit meneur de jeu ajaccien c'était à une bonne trentaine de mètres et euh **il y a quelques instants y a cinq six minutes environ** c'est Mickaël Pagis de l'autre côté côté sochalien qui lui aussi avait tenté sa chance c'était à trente mètres également et **là** ça a failli rentrer puisque la frappe du droit terrible de Mickaël Pagis est venue titiller la barre transversale d'Hervé Sekli Mickaël Pagis qui se sent ici un petit peu chez lui [...] depuis euh eh bien on va dire que les Sochaliens ont bien rétabli euh la balance au niveau du jeu les Ajacciens ont largement dominé **les dix premières minutes** mais les Sochaliens **maintenant** arrivent à se montrer dangereux notamment »

• « et **une petite minute de moins** ici à Bordeaux **vingt-quatre minutes de jeu** toujours zéro à zéro avec une très belle occasion **il y a quelques instants** pour les Girondins une balle parfaitement distillée longue en profondeur dans l'axe lensois lancée par euh Christophe Dugarry à l'intention et à l'attention de Pauleta »

En ligne chaque intervention est précédée de l'indication du temps de jeu écoulé et les interventions sont ordonnées en fonction de cette valeur, même en cas d'écriture d'une intervention en différé après saisie d'autres informations (comme nous l'avons vu précédemment, § 2., avec les informations de 17h43 remplacées par celles de 18h12).

3.2.3 Indications de localisation

Comme les indications temporelles, les indications de localisation peuvent être absolues dans les deux corpus ou relatives à la position de l'énonciateur (avec *ici* et, surtout, *là-bas*) dans les multiplex seulement (comme dans les extraits suivants).

Indications de localisation absolue (**en gras**) et relatives à la position de l'énonciateur (**en italique gras**) dans des interventions du Corpus footballistique de multiplex :

- « [...] toujours zéro zéro et on joue depuis maintenant trente et une minute **ici** à Bollaert »
- « vingt-trois minutes de jeu toujours un but à zéro pour le Stade Rennais le match est toujours aussi engagé et plaisant les Rennais à l'attaque avec Olivier Monterrubio **là-bas sur le côté droit** Monterrubio le ballon sur son pied gauche il essaye de trouver Piquionne le buteur [...] »

Pour les diverses raisons évoquées, nous pouvons conclure que, selon le média de diffusion, les commentaires qui permettent au public de suivre les matchs qu'il ne voit pas ne l'informent pas de manière équivalente, et donc qu'ils ne fournissent pas exactement la même matière une fois compilés en corpus.

4 Chercher à améliorer la qualité des extractions en balisant plus finement les corpus

Sans avoir épuisé l'étude de ce qui rapproche et différencie les deux corpus constitués, concentrons-nous maintenant sur ce qui pourrait être introduit dans le balisage XML afin d'améliorer le pouvoir documentaire des extractions faites.

La qualité informationnelle des corpus considérés est à envisager en fonction de ce pour quoi ils sont explorés. Il s'agit dans ce contexte de la documentation de lexicographes pour la description d'unités lexicales à traiter dans un dictionnaire spécialisé ou général qui prendrait en compte les emplois de ces items dans des commentaires footballistiques.

Un exemple de description spécialisée minutieuse qui peut être visée en s'appuyant sur le corpus a été proposé dans (Gasiglia 2012).

Ce qui est envisagé est donc diversement rapprochable des travaux éditoriaux et scientifiques conçus dans des perspectives comparables aux nôtres :

- des dictionnaires spécialisés comme Blanchet & Lesay (2011), Lesay (2006), Schmidt (*Kicktionary*, en ligne) ;
- des travaux linguistiques comme Deulofeu (2000), Galisson (1978), Gross & Guenther (2002), Lavric, Pisek, Skinner & Stadler eds (2008), Leroyer & Møller (2004), Schmidt (2006 ; 2007 ; 2008a ; 2008b ; 2009a ; 2009b ; 2010), Song (2003) ;

ou dans d'autres perspectives, comme cela est le cas pour d'autres dictionnaires spécialisés comme Bouchard (1996), Doillon (2002), Ligas (2008), Merle (1998), Meyer (2012), Montvalon (1998), Perret (2002), Petiot (1982), Vandel (1992).

Si les deux corpus ont à l'heure actuelle un haut rendement exploratoire du fait de leur thématisation et de la sélection de commentaires, donc d'énoncés de locuteurs spécialisés qui s'adressent à un large public, la spécificité des situations d'énonciation propres aux interventions de chaque corpus a un impact sur la nature des données observables au sein de chacun (cf. *supra*). Les observations faites invitent donc à envisager d'affiner le balisage de chacun des corpus.

Les enrichissements considérés peuvent prolonger les repérages d'entités nommées qui ont été mis en œuvre à partir des listes de noms de joueurs, par exemple, ce qui permettrait d'identifier les référents des pronoms ou des syntagmes qui dénotent les mêmes entités. Ils peuvent par ailleurs résulter d'un processus d'analyse, comme cela pourrait se concevoir pour l'annotation des constructions syntaxico-sémantiques de certains prédicats. Un parallèle peut être fait sur ce dernier point avec les exemples que Thomas Schmidt a extraits d'un autre corpus footballistique et présentés, dans le *Kicktionary* – dictionnaire trilingue (anglais / allemand / français) du football en ligne –, enrichis d'annotations résultant d'analyses faites dans le cadre de la Frame Semantics :

• Présentation du corpus documentaire : « The core corpus used in this project is a collection of German, English and French football match reports from the UEFA website (www.uefa.com). For each language, there are roughly 500 such texts, amounting to around 200,000 words. This core corpus is partly parallel, i.e. many texts (about half of them) are direct translations of one another. For German, the corpus contains additional match reports from the website of the football journal kicker (www.kicker.de - about 1,200 texts = 750,000 words) and about an hour (= ca. 10,000 words) of transcribed spoken soccer commentary from German radio. » (cf. <http://www.kicktionary.de/background.html> § Corpus)

• « Éléments de Frame » repérés comme étant saillants dans les énoncés annotés retenus comme « Exemples » pour le verbe *dégager* :

– « **Éléments de Frame**

BALL [Ball] GOALKEEPER [Goalkeeper] INTERVENING_PLAYER [Player] INTERVENTION_LOCATION [On_The_Field_Location] PART_OF_BODY [Part_Of_Body] SHOT [Moving_Ball] »

– « **Exemples**

1. [...] [Aleksei Berezoutski]_{INTERVENING_PLAYER} intervenait pour **dégager** [le ballon]_{BALL}. [75243 / p5]
2. [Ball]_{INTERVENING_PLAYER} **finissait** tout de même **par dégager** [la balle]_{BALL} mais Barry van Galen, à la récupération, [...]. [79758 / p8]
3. [Luis García]_{GOALKEEPER} n'**arrivait pas à dégager** [un coup-franc de Libor Sionko]_{SHOT} [...]. [80118 / p3]
4. [...] Mark Schwarzer bloquait la balle de la main gauche avant de [la]_{BALL} **dégager** [du pied gauche]_{PART_OF_BODY}. [79748 / p5]
5. [...] [Costinha]_{INTERVENING_PLAYER} **dégageait** [la tête de Raio Piiroja]_{SHOT} [sur la ligne]_{INTERVENTION_LOCATION}. [75228 / p5]
6. [...] [Fernando Couto]_{INTERVENING_PLAYER} **dégageait** [son lob]_{SHOT} [de la tête]_{PART_OF_BODY}. [79740 / p7] »

Je ne reviendrai pas ici sur l'annotation des pronoms et syntagmes dénotant des entités : elle constitue un travail minutieux mais ne pose pas de problèmes particuliers.

L'annotation des structures prédicatives, elle, implique déjà que l'expression de chaque prédicat soit délimitée. C'est cette question parfois complexe que je vais aborder maintenant en revenant sur le matériau linguistique utilisé pour narrer les actions de jeu.

Alors qu'à l'oral les énoncés peuvent être longs, pas toujours bien structurés et très riches en propositions (en particulier relatives) imbriquées ou concaténées (comme dans l'exemple ci-après pour les multiplex ou comme le décrit Deulofeu (2000) pour un commentaire télévisuel), dans le corpus écrit de *lequipe.fr* les phrases sont complètes et plutôt courtes.

« en regardant les Girondins après vingt minutes de jeu en seconde mi-temps des Girondins qui n'arrivent pas du tout toujours zéro à zéro bien sûr les Girondins qui n'arrivent pas du tout à accélérer le cours du jeu qui baissent même le pied on dirait qu'y a un coup de pompe physique et à ce petit jeu à ce jeu qui devient petit eh bien ce sont les Lenois qui font la bonne opération qui sont parfaitement regroupés sur leur solide défense qui procèdent par contre-attaque »

Ceci semble (au moins en partie) induit par le fait qu'à la radio le reporter narre une action pour laquelle il sélectionne au fil de son énoncé ce qui semble être important afin de rendre précisément compte de ce qui

se passe, alors qu'en ligne le commentateur connaît l'issue de l'action et peut ne retenir que ce qui s'est avéré être le plus crucial. Observons les données de manière plus détaillée.

4.1 Phrases simples

Les interventions du Web comportent en moyenne deux phrases, qui sont plutôt des phrases simples :

– Ces dernières sont généralement des phrases à un ou, éventuellement, deux prédicats verbaux coordonnés par *et* ou par *mais*.

– Les phrases nominales sont, elles aussi, bien représentées mais il arrive qu'elles se limitent à un mot comme *Corner*, ce qui peut alors être redondant avec le pictogramme qui précède l'énoncé.

– Par ailleurs un petit nombre de phrases indépendantes adverbiales à sens locatif (comme *Assez nettement à côté.*) ou prépositionnelles à sens évaluatif (comme *Sans danger pour les visiteurs.*) s'observe.

En fonction de ces observations, il apparaît que le balisage des segments relatifs à la narration des différentes actions de jeu est raisonnablement faisable dans ces phrases simples du corpus Web, de même qu'il est possible de caractériser la nature de leurs têtes lexicales.

Dans le corpus de multiplex par contre, si les phrases simples ne sont pas rares, d'une part l'intonation en associe couramment plusieurs en séquences dont la délimitation est nécessairement dépendante des interprétations de celui qui analyse le corpus, et d'autre part il y a des procédés typiques de l'oral qui cassent le découpage canonique des phrases en prédicats et arguments.

Le premier exemple suivant fournit un contexte plutôt favorable à l'isolement des phrases nominales dont les têtes sont les deuxième et troisième occurrences de *corner*, alors que le second est un exemple d'entrelacs propositionnel, avec incise de *c'est vrai* et connexion par *ce qui fait que*, dont l'analyse peut être source de variations.

- « [...] attention Cyril Chapuis dans la surface et Viata qui réussit à revenir à mettre le pied droit et à mettre ce ballon en corner **corner** nouveau **corner** pour les Marseillais qui mènent [...] »
- « toujours zéro à zéro toujours agréable mais les deux équipes **c'est vrai** avec leur talent se marquent beaucoup à la culotte **ce qui fait qu'**on n'a pas d'ouverture de score après quarante et une minutes de jeu ici à Bordeaux »

4.2 Phrases complexes

Les rares propositions infinitives attestées dans les deux corpus (comme ci-après, dans un multiplex) ont des valeurs causales, finales ou résultatives :

- « [...] Hector Tapia à l'affût dans la surface de réparation **pour reprendre ce ballon de la tête** une tête piquée directement dans la cage de Ronan Le Crom [...] »

Les propositions relatives – essentiellement introduites par *qui* (comme on le voit ci-dessous) – servent à exprimer des enchaînements d'actions mais, alors qu'elles sont omniprésentes à l'oral, elles s'observent moins fréquemment dans les phrases écrites, dans lesquelles elles jouent cependant un rôle important pour l'articulation temporelle des procès.

- Dans le Corpus footballistique de multiplex :
« et après six minutes toujours euh zéro à zéro mais nette domination des joueurs de Créteil **qui** se sont procuré deux occasions très franches notamment par l'intermédiaire de ce jeune russe Kossonogov prêté par les Girondins de Bordeaux à Créteil et **qui** a tout à l'heure devancé la sortie de Tingry à la limite de la surface de réparation et d'une pichenette eh bien son ballon est passé tout près des buts **qui** étaient donc vides et quatre minutes plus tard ce même Kossonogov eh bien a donné un ballon en or à Sébastien Dallet le capitaine de Créteil **qui** se présente tout seul face à Tingry et cette fois euh Tingry sauve son camp en expédiant ce ballon en corner côté rémois

une seule petite occasion un tir en pivot d'Olivier Picqueu de vingt mètres **qui** est passé quand même à côté des buts de Legrand donc pour l'instant zéro zéro »

- Dans le Corpus footballistique Web (*lequipe.fr*) au fil du match :
 - « [...] Anin et Richert s'y prennent à deux pour dégager le ballon devant Marchal **qui** s'était élevé assez haut. »
 - « [...] Guié Guié temporise et décale vers Meriem. Celui-ci trouve Hellebuyck, **qui** perfore la défense bordelaise et finit habilement du pied gauche. »

4.3 Syntagmes participiaux

En outre, des syntagmes, comme ceux ci-dessous – dont la tête est un participe passé, qui sont apposés à des sujets de verbes finis et qui dénotent un procès antérieur à celui exprimé par le verbe – s'observent couramment dans le corpus issu du Web.

« *Servi dans le dos de la défense par Pedretti, Sow se présente seul face à Gabulov.* »

Ils sont par contre rarement attestés dans les multiplex : sur les huit occurrences de la forme *servi*, on en a deux seulement dans des conditions comparables, et six dans des relatives avec un auxiliaire :

- « *attention avec ce contre euh monégasque avec Ludovic Gallardo **servi** sur le côté droit qui va tenter la reprise* »
- « *on les a vus tout à l'heure avec un nouveau euh déboulé de Giuly sur euh le côté droit **qui avait été servi** admirablement par euh Gallardo* »

4.4 Syntagmes nominaux

Les prédicats nominaux associés à des verbes complètent enfin le matériau utilisé pour articuler les mentions d'actions de jeu liées, que le procès ou l'événement exprimé par la proposition nominale soit un argument du verbe, ou qu'il situe dans le temps celui exprimé par le verbe :

Prédicats nominaux (**en gras**) et verbaux articulés dans le Corpus footballistique de multiplex afin d'exprimer des enchaînements d'actions – les parenthèses délimitent les descriptions d'actions et les chiffres qui leur sont postposés indiquent l'ordre des actions :

- « [...] (une **reprise** de volée extraordinaire en ciseaux (sur un **centre** de Fadiga)₁ permet à l'AJ Auxerre de reprendre l'avantage)₂ [...] »
- « [...] déjà à la neuvième minute les Montpelliérains s'étaient montrés dangereux (sur un **corner** de Barbosa)₁ ((une **dévi**ation de la tête de Sylvestre)₂ et (une **reprise** du pied gauche au deuxième poteau de Mansare)₃) avaient obligé Grégorini à dévier en corner)₄ sinon c'est une grosse domination niçoise [...] »
- « [...] Sochaux (qui a été là à deux doigts d'égaliser)₂ (après un joli **contre**)₁ [...] »

Comme les propositions relatives, les prédicats nominaux sont moins nombreux à l'écrit qu'à l'oral, mais ils s'y observent néanmoins régulièrement.

Certains des extraits de corpus présentés dans cet article rendent perceptible le fait que les combinaisons de prédicats permettent de construire des énoncés complexes au sein desquels la délimitation de chaque description d'action peut parfois être malaisée.

Si cette délimitation était néanmoins opérée, le balisage mis en place pourrait utilement marquer les successions d'actions (comme cela a été fait dans les exemples qui viennent d'être présentés) et leurs éventuelles relations, en particulier causales.

Ce type d'enrichissement permettrait de très sensiblement améliorer la qualité informationnelle des corpus pour des extractions semi-automatiques comme pour des environnements de lecture enrichis destinés à des utilisateurs ayant du mal à comprendre les subtilités du jeu et donc des actions décrites, malgré leur effort documentaire.

4.5 Reprises

Terminons notre réflexion sur le balisage des actions de jeu par un rapide retour sur une observation déjà évoquée : une même action ou séquence de jeu peut être mentionnée plusieurs fois, soit au sein d'une même intervention (comme cela est illustré dans les premiers exemples sortis du § 3.2), soit au sein de plusieurs interventions relatives au même match (il s'agit alors d'un rappel), soit encore lors d'autres matchs (ce qui ne s'observe que si ce qui est évoqué est notoirement important).

Le corpus de multiplex offre des exemples des trois types, alors que celui issu du Web, cf. ci-après, en contient peu et plutôt du second type :

- Dans le cours du match, puis à la fin de celui-ci, la même action est décrite, mais de manières différentes :
 - « [88^e minute] Sur un long coup franc de Batlles le long de la ligne de touche côté gauche, Clément dévie de la tête au second poteau pour **Marchal qui a surgi à temps**. Saint-Etienne se libère et c'est mérité ! »
 - « [90^e+4 minutes] Grâce à un **but du défenseur Sylvain Marchal** à deux minutes de la fin, Saint-Etienne arrache une victoire méritée. Face à des Lorrains qui étaient venus chercher le nul, les Verts ont débloqué une situation mal embarquée et prennent provisoirement la tête. »
- Dans deux interventions consécutives au cours du match, l'action narrée dans la première est reprise par la mention de son agent dans la seconde :
 - « [82^e minute] Olympiakos ouvre le score. À la réception d'une belle passe de Yeste par-dessus la défense marseillaise, **Fetfatzidis, côté gauche dans la surface, reprend d'une volée limpide du gauche**. Le ballon termine sa course dans le petit filet gauche de Mandanda. »
 - « [84^e minute] D'après le ralenti, **le buteur** était en position de hors-jeu au départ de l'action... »

Dans le cadre d'un processus de balisage des descriptions d'actions de jeu, le fait que différents segments d'énoncés narrent les mêmes actions mériterait d'être marqué : la première mention d'une action servant de repère associé à un identifiant auquel les autres devraient référer (au moyen d'un attribut IDREF).

5 Conclusion

Dans le cadre de cette contribution, mon objectif était de présenter une sélection de données caractéristiques des ressemblances et différences observées entre les deux corpus de commentaires footballistiques, afin de dresser un inventaire de ce qui mériterait d'être annoté au sein des différentes interventions des commentateurs, puis de m'interroger sur les modalités de mise en œuvre de ces annotations.

Je me suis concentrée sur les délimitations des descriptions d'actions de jeu afin de valoriser les traits saillants de leur matériau discursif en fonction du type de média.

La complexité des données fait que le balisage envisagé risque de susciter certaines difficultés et donc mériterait d'être réalisé parallèlement par plusieurs annotateurs¹⁴, afin de confronter leurs décisions et de revenir sur leurs points de désaccord.

Références bibliographiques

- Blanchet, B. & Lesay, J. D. (2011). *Le dico du parler sport*. Paris : Fetjaine.
- Bouchard, J.-P. (1996). *Les mots du sport. La tête dans le guidon*. Paris : Éditions du Seuil.
- British National Corpus*. <http://www.natcorp.ox.ac.uk/corpus/>.
- Collins COBUILD English Language Dictionary*. London / Glasgow : Collins. 1987.
- Corbin, P. (2005). Des occurrences discursives aux contextualisations dictionnaires. Éléments d'une recherche en cours sur l'expression en français d'expériences du football. In M. Heinz (éd.). *L'exemple lexicographique dans*

- les dictionnaires français contemporains. Actes des "Premières Journées allemandes des dictionnaires" (Klingenberg am Main, 25-27 juin 2004).* Lexicographica series maior, vol. 128. Tübingen : Max Niemeyer Verlag. 125-156.
- Corbin, P. (2008). Peut-on parler d'une langue du football ? Réflexions sur une expérience de constitution et d'exploitation d'une ressource discursive informatisée. In F. Maniez, P. Dury, N. Arlin & C. Rougemont (éds). *Corpus et dictionnaires de langues de spécialité*. Grenoble : Presses Universitaires de Grenoble. 271-300.
- Deulofeu, J. (2000). Les commentaires sportifs télévisés sont-ils un genre au sens de la "grammaire des genres" ? In M. Bilger (éd.). *Corpus. Méthodologie et applications linguistiques*. Paris : Honoré Champion Éditeur / Perpignan : Les Presses Universitaires de Perpignan. 271-295.
- Doillon, A. (2002). *Le dico du sport*. Paris : Fayard.
- Fort, K. & Claveau, V. (2012a). Annotation manuelle de matchs de foot : Oh la la ! l'accord inter-annotateurs ! et c'est le but ! In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, Grenoble, 4-8 juin 2012*. Vol. 2. 383-390. <http://aclweb.org/anthology-new/F/F12/F12-2031.pdf>.
- Fort, K. & Claveau, V. (2012b). Annotating football matches: influence of the source medium on manual annotation. In *LREC 2012 Proceedings, Istanbul, 21-27 May 2012*. 2567-2572. <http://www.lrec-conf.org/proceedings/lrec2012/index.html>.
- 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.
- Gasiglia, N. (2004). Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus. *Revue française de linguistique appliquée, IX.1*, 45-62.
- Gasiglia, N. (2005). Stratégie de constitution de corpus oraux transcrits (1) : arguments pour un corpus plurithématique à haut rendement. In G. Williams (dir.). *La linguistique de corpus en France ou en français*. Coll. Rivages linguistiques. Rennes : Presses Universitaires de Rennes. 219-232.
- Gasiglia, N. (2008). Stratégie de consultation de corpus oraux transcrits : pistes méthodologiques pour l'exploration d'un corpus thématique à haut rendement. In G. Williams (éd.). *Actes des Troisièmes Journées de la Linguistique de Corpus*. *Revue électronique Texte et Corpus*, 145-164. http://web.univ-ubs.fr/corpus/jlc3/2_5_gasiglia.pdf.
- Gasiglia, N. (2010). *Des usages en corpus aux descriptions dictionnairiques*. Habilitation à diriger des recherches. 3 vol. Université Lille 3.
- Gasiglia, N. (2012). Exploiter les énoncés de corpus thématiques à haut rendement pour décrire des prédicats verbaux de lexiques spécialisés de masse. *Lexicographica, 28*, 207-232.
- Gross, G. & Guenther, F. (2002). Comment décrire une langue de spécialité ? *Cahiers de lexicologie, 80*, 179-199.
- Kicktionary. The multilingual electronic dictionary of football language*. <http://www.kicktionary.de/>.
- Kilgarrieff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29.3*, 333-348. Nouv. éd. In T. Fontenelle (ed.) (2008). *Practical Lexicography. A reader*. Oxford : Oxford University Press. 89-101.
- Lavric, E., Pisek, G., Skinner, A. & Stadler, W. (eds) (2008). *The Linguistics of Football*. Tübingen : Gunter Narr Verlag.
- Le Dictionnaire Hachette-Oxford français-anglais / anglais-français*. Oxford, New York, Toronto : Oxford University Press / Paris : Hachette Livre. 1994.
- Leroyer, P. & Møller, B. (2004). Les nouveaux habits de la lexicographie spécialisée : intégration de la métaphorique dans le dictionnaire du football. In G. Williams & S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient, France. July 6-10, 2004*. Lorient : Université de Bretagne-Sud. Vol. II. 571-582.
- Lesay, J. D. (2006). *Les mots du football*. Paris : Belin.
- Ligas, P. (2008). *Dictionnaire alphabétique et analogique du français des activités physiques et sportives*. 2 vol. Verona / Bolzano : QuiEdit.
- Merle, P. (1998). *L'argot du foot*. Paris : Mona Lisait.
- Meyer, B. (2012). *Dictionnaire du football. Le ballon rond dans tous ses sens*. Paris : Honoré Champion Éditeur.

- Montvalon, C. de (1998). *Le dico du foot*. La Tour d'Aigues : Éditions de l'Aube.
- Perret, P. (2002). *Le parler des métiers. Dictionnaire thématique alphabétique*. Paris : Robert Laffont.
- Petiot, G. (1982). *Le Robert des sports. Dictionnaire de la langue des sports*. Paris : Le Robert.
- Schmidt, T. (2006). Interfacing lexical and ontological information in a multilingual soccer FrameNet. In *Proceedings of OntoLex 2006. Interfacing ontologies and lexical resources for semantic Web technologies, Genoa, Italy, May, 24-26, 2006*. Paris : ELRA. <http://www.kicktionary.de/REOURCES/schmidt2006.pdf>.
- Schmidt, T. (2007). The Kicktionary: a multilingual resource of the language of football. In G. Rehm, L. Lemnitzer & A. Witt (eds). *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen / Data Structures for Linguistic Resources and Applications. Proceedings der GLDV Frühjahrstagung 2007*. Tübingen : Gunter Narr Verlag. 189-196.
- Schmidt, T. (2008a). The Kicktionary: combining corpus linguistics and lexical semantics for a multilingual football dictionary. In E. Lavric, G. Pisek, A. Skinner & W. Stadler (eds). 11-23.
- Schmidt, T. (2008b). The Kicktionary revisited. In A. Storrer, A. Geyken, A. Siebert & K.-M. Würzner (eds). *Text Resources and Lexical Knowledge. Selected papers from the 9th Conference on Natural Language Processing, KONVENS, 2008*. Berlin / New York : Mouton de Gruyter. 239-252.
- Schmidt, T. (2009a). The Kicktionary – A multilingual lexical resource of football language. In H. C. Boas ed. *Multilingual Framenets in Computational Lexicography. Methods and applications*. New York : Mouton de Gruyter. 101-134.
- Schmidt, T. (2009b). Kicktionary. In P. Schlobinski & A. Burkhard (eds). *Flickflack, Foul und Tsukahara: der Sport und seine Sprache*. Mannheim : Duden Verlag. 117-132.
- Schmidt, T. (2010). Der Fußballwortschatz im Kicktionary. In *Der Deutschunterricht* 3/10 (« Fußball und Sprache »). 17-25.
- Sinclair, J. (1996). *Preliminary Recommendations on Corpus Typology*. Rapport technique. EAGLES (Expert Advisory Group on Language Engineering Standards). Mai 1996.
- Song, C.-M. (2003). *Rôles et parcours actantiels dans les sports collectifs : le cas du football. Contribution à une sémiotique des pratiques sportives*. Thèse de doctorat. Université de Limoges.
- TEI: P5. 2007. <http://www.tei-c.org/Guidelines/P5/>.
- Vandel, P. (1992). *Le dico français / français*. Paris : JC Lattès.

¹ Merci aux relecteurs d'une version préliminaire de cette contribution pour les suggestions formulées.

² Plus précisément l'actuel parcours LTTAC du master mention Sciences du langage de l'Université Lille 3 (cf. <http://perso.univ-lille3.fr/~ngasiglia/M.LTTAC/>), que nous avons animé Pierre Corbin et moi de 1999 à 2011 et que j'anime seule depuis.

³ Cf. Sinclair (1996), et le *British National Corpus*, constitué entre 1991 et 1994.

⁴ Cf. le *Collins COBUILD English Language Dictionary* (1987) et *Le Dictionnaire Hachette-Oxford français-anglais / anglais-français* (1994), pour n'évoquer que les premiers dictionnaires monolingue et bilingue français-anglais exploitant un corpus électronique.

⁵ Cf. Kilgarriff & Grefenstette (2003) ou le workshop « Web as Corpus » (<http://www.aclweb.org/anthology/W/W06/W06-1700.pdf>) organisé par Adam Kilgarriff & Marco Baroni en 2006, et le Sketch Engine développé par Adam Kilgarriff (<http://www.sketchengine.co.uk/>).

⁶ Ce point a été développé dans plusieurs de nos publications : Corbin (2005 ; 2008) et Gasiglia (2004 ; 2005 ; 2008 ; 2010 ; 2012).

⁷ Les étudiants de la promotion 2002-2003 du DESS Lexicographie et terminographie (prédécesseur du master 2^e année LTTAC) ont transcrit et relu les commentaires enregistrés avant que nous fassions, Pierre Corbin et moi, de nouvelles relectures minutieuses. Des consignes de transcription et de balisage avaient été établies avec les étudiants transcripateurs afin que leurs productions soient aussi cohérentes que possible, mais nos relectures se sont avérées

utiles pour l'homogénéisation (encore imparfaite) des choix de différentes natures (interprétation du flux audio, variantes graphiques, balisage de certains objets annotés).

⁸ Les noms de joueurs saisis étaient comparés à ceux des feuilles de matchs (les constitutions d'équipes) fournies par le site *lequipe.fr* pour chaque match et à ceux figurant dans les pages décrivant, sur le même site, les carrières des joueurs. Les noms de clubs, villes et stades avaient été inventoriés, mais nous n'avions relevé que ceux qui semblaient *a priori* utiles, en fonction des matchs transcrits, ce qui s'est révélé insuffisant car les commentateurs évoquaient aussi des entités non directement liées aux matchs commentés. Les noms d'entités qui ne faisaient pas partie des inventaires devaient être repérés et annotés sans pouvoir se référer à un index préalablement constitué, mais ils ont fait l'objet de contrôles *a posteriori*.

⁹ Cf. <https://www.kuleuven-kulak.be/~hpauluss/>.

¹⁰ Cf. <http://macaon.lif.univ-mrs.fr/>.

¹¹ Bien que les commentaires des matchs diffusés à la télévision semblent souvent être moins précisément commentés oralement qu'ils ne le sont à la radio, probablement parce que les téléspectateurs peuvent voir les actions se dérouler, la constitution d'un corpus vidéo permettrait de mieux étudier notamment les expressions de localisations sur les terrains. Afin de profiter des images tout en se donnant les moyens de comparer les types de commentaires, il serait intéressant de se doter de commentaires radiophoniques, saisis en direct sur le site *lequipe.fr* et télévisuels des mêmes matchs, et de synchroniser les textes de chacun sur la base du temps de jeu écoulé.

¹² Les inventaires de noms d'entités élaborés en 2002-2003 (cf. n. 8) ont, à cette fin, été actualisés.

¹³ Le second élément <p> de chaque <div> ne contient qu'un espace insécable. Il pourrait être supprimé et son attribut @type, qui indique le nom du pictogramme affiché en ligne, transféré sur l'élément <p> suivant.

¹⁴ Fort & Claveau (2012a ; 2012b), eux, procèdent de même afin d'analyser les accords inter- et intra-annotateurs.