Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but !

Karen Fort, Vincent Claveau

▶ **To cite this version:**

Karen Fort, Vincent Claveau. Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but !. TALN - Traitement Automatique des Langues Naturelles, Jun 2012, Grenoble, France. pp.383-390. ⟨hal-00709181⟩

HAL Id: hal-00709181

https://hal.science/hal-00709181v1

Submitted on 18 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Annotation manuelle de matchs de foot : Oh la la ! l'accord inter-annotateurs ! et c'est le but !

Karën Fort[1, 2] Vincent Claveau[3]
(1) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy
(2) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430 Villetaneuse
(3) IRISA - CNRS, Campus de Beaulieu, 35200 Rennes
karen.fort@inist.fr, vincent.claveau@irisa.fr

RÉSUMÉ

Cet article présente une campagne d'annotation de commentaires de matchs de football en français. L'annotation a été réalisée à partir d'un corpus très hétérogène, contenant à la fois des comptes-rendus minute par minute et des transcriptions des commentaires vidéo. Nous montrons ici comment les accords intra- et inter-annotateurs peuvent être utilisés efficacement, en en proposant une définition adaptée à notre type de tâche et en mettant en exergue l'importance de certaines bonnes pratiques concernant leur utilisation. Nous montrons également comment certains indices collectés à l'aide d'outils statistiques simples peuvent être utilisés pour indiquer des pistes de corrections des annotations. Ces différentes propositions nous permettent par ailleurs d'évaluer l'impact des modalités sources de nos textes (oral ou écrit) sur le coût et la qualité des annotations.

ABSTRACT

Manual Annotation of Football Matches : Inter-annotator Agreement ! Goooooal !

We present here an annotation campaign of commentaries of football matches in French. The annotation was done from a very heterogeneous text corpus of both match minutes and video commentary transcripts. We show how the intra- and inter-annotator agreement can be used efficiently during the whole campaign by proposing a definition of the markables suited to our type of task, as well as emphasizing the importance of using it appropriately. We also show how some clues, collected through statistical analyses, could be used to help correcting the annotations. These statistical analyses are then used to assess the impact of the source modality (written or spoken) on the cost and quality of the annotation process.

MOTS-CLÉS : annotation manuelle, accords inter-annotateurs.

KEYWORDS: manual annotation, inter-annotator agreement.

1 Introduction

Nous étudions dans cet article la création d'un corpus textuel annoté construit à partir de transcriptions de commentaires vidéos et de sites Web spécialisés. Ce corpus annoté est développé dans le but de mettre au point des techniques automatiques d'analyse, tels que le résumé vidéo, le *repurposing* (transformation du contenu et du format pour un autre support de diffusion) ou l'extraction d'information pour les vidéos d'événements sportifs. Cette application, développée dans le cadre d'un partenariat industriel, n'est pas détaillée plus avant dans cet article, mais il est important de noter qu'elle guide la définition des éléments à annoter (cf. section 2).

Outre la présentation d'une nouvelle ressource annotée, cet article a pour objectif de montrer l'intérêt d'analyses fines pour évaluer la qualité d'une telle ressource hétérogène. En particulier, nous proposons une définition des mesures d'accord inter- et intra-annotateur adaptée à ce type d'annotation où seuls certains éléments des corpus sont annotés. Nous montrons également comment certains indices collectés à l'aide d'outils statistiques simples peuvent être utilisés pour souligner les difficultés de la tâche d'annotation et indiquer des pistes de corrections des annotations. Ces différentes propositions nous permettent par ailleurs d'évaluer l'impact des modalités sources de nos textes (oral ou écrit) sur le coût et la qualité des annotations.

D'un point de vue applicatif, quelques travaux (Nemrava *et al.*, 2007, par exemple) font référence à un corpus annoté du domaine du football, mais à notre connaissance, aucun ne détaille l'annotation du corpus utilisé. D'autres études ont fait usage de corpus de football pour créer des lexiques monolingues (Gasiglia, 2003) or multilingues (Schmidt, 2008) plus ou moins détaillés. Dans ces cas, si les publications associées détaillent l'annotation du corpus utilisé, les annotations elles-même sont de nature linguistique plutôt que du domaine et soulèvent des questions différentes. D'un point de vue méthodologique, l'analyse statistique des annotations repose principalement sur les calculs d'accord inter-annotateurs (Artstein et Poesio, 2008, pour une revue détaillée). Ces derniers sont généralement fournis sur les corpus annotés comme mesure d'évaluation de la qualité de la ressource produite (Dandapat *et al.*, 2009, *inter alia*). Les méthodes d'annotation agiles (Voormann et Gut, 2008) proposent d'utiliser ces mesures pendant toute l'annotation du corpus, pour assurer la cohérence des annotations et limiter les divergences dans les cas, majoritaires, où l'on ne peut pas tout annoter en double avec adjudication. Notre travail se situe dans ce cadre mais aborde plusieurs problèmes posés par les particularités de nos annotations. Après une présentation des données et des annotations en section 2, nous détaillons les différentes analyses menées en section 3 et nous concluons en donnant quelques perspectives à ce travail.

2 Campagne d'annotation

2.1 Données, annotations et méthodologie

Le corpus annoté couvre 16 matchs de football. Il est composé de 24 transcriptions de commentaires tirés de vidéos (1 par mi-temps, 12 matchs) et de 16 fichiers contenant une description minute-par-minute du match (dont les 12 de la transcription et 4 matchs additionnels) tirés de sites Web spécialisés. La parole contenue dans les vidéos a été transcrite manuellement en utilisant TRANSCRIBER (Barras *et al.*, 1998) et son guide de transcription par défaut. L'ensemble du corpus a une taille d'environ 250 000 mots. Sa principale caractéristique est d'être très hétérogène

(Fort *et al.*, 2011), que ce soit d'un point de vue des types de match (ligues, championnats...), de la taille des fichiers (de 1 116 tokens par match pour les minutes à 21 000 tokens pour les transcriptions), ou de la source (chaînes de diffusion des vidéos, commentateurs, sites Web...).

Le jeu d'étiquettes a été construit en définissant les éléments intéressants pour l'application finale et ensuite affiné durant les phases d'entraînement et de pré-campagne. L'ensemble des étiquettes retenues a été divisé en trois couches, *Unités*, *Actions* et *Relations* (cf. tableau 1 [1]), chacune correspondant à un niveau d'analyse de complexité croissante à aborder successivement par les annotateurs. Par cohérence avec les besoins applicatifs et pour prendre en compte le style elliptique de l'oral (« `Makoun. Et c'est récupéré. Clerc, avec Cris. Boumsong, Makoun.` »), nous avons décidé de ne pas faire porter les annotations sur les prédicats dénotant les actions ou les relations, souvent absents, mais sur les acteurs impliqués.

Unités	acteurs	*Joueur, Equipe, Arbitre, Entraineur, ArbitreAssistant, Président*
	circonstants	*EspaceSurTerrain, LieuDuMatch, TempsDansMatch*
Actions	arbitrales	*TirerCoupFrancDirect, TirerCoupFrancIndirect, TirerCorner, TirerPenalty, FaireFauteDeJeu, HorsJeu, MarquerBut, PrendreCartonJaune, PrendreCartonRouge, PrendreRappelALOrdre*
	autres	*Centrer, FaireTentative2Centre, Dribbler, RaterBut, ArreterBut, IntercepterBallon, PossederBallon, ActionDuPublic*
Relations	arbitrales	*FaireFauteSurJoueur, TaclerFaute, RemplacerJoueur*
	autres	*FaireCombinaison, FairePasse, FaireTentative2Passe*

TABLE 1 – Couches d'annotations retenues et étiquettes correspondantes

La méthodologie employée pour l'annotation de ce corpus suit les recommandations de Bonneau-Maynard *et al.* (2005) et Gut et Bayerl (2004) ; elle est décrite en détail dans (Fort et Claveau, 2012). L'annotation a été réalisée par deux annotateurs experts du domaine avec l'outil d'annotation GLOZZ (Widlöcher et Mathet, 2009), choisi en raison de sa facilité d'utilisation et de la possibilité qu'il offre d'annoter des relations. Les temps d'annotation par couche ont été mesurées à l'aide de l'outil TIMETRACKER [2]. Nous avons également invité les annotateurs à ajouter des commentaires sur leurs annotations, et un attribut *Incertitude* a été mis à leur disposition dans GLOZZ.

2.2 Données générales sur le processus d'annotation

Le nombre total d'annotations produites s'élève à 37 784 dont 27 736 (soit plus de 73 %) pour les transcriptions. Toutes les catégories ont été utilisées, mais avec une grande disparité : par exemple, *TirerCoupFrancIndirect* et *TirerPenalty* n'ont servi que 2 fois (et uniquement dans les minutes), *PrendreCartonRouge* 6 fois et *Président* 9 fois.

Le tableau 2 présente le temps d'annotation moyen (pour 1 000 tokens) par annotateur et par source. Un t-test de Welsh à deux échantillons (avec $p = 0,05$) montre que les différences entre annotateurs ne sont pas significatives, que ce soit pour les transcriptions ou pour les

1. Le regroupement des étiquettes à l'intérieur de ces couches (circonstants, acteurs, etc) est proposé ici pour faciliter la lecture et l'analyse, mais n'existait pas dans le modèle de données utilisé pour l'annotation.

2. `http://www.formassembly.com/time-tracker/#`

	Minutes	Transcriptions
Annotateur 1	36,92	20,03
Annotateur 2	41,30	16,06

TABLE 2 – Temps moyen d'annotation par source et par annotateur, en minute/1 000 tokens

minutes. En revanche, les différences entre modalités sont jugées statistiquement significatives, pour les deux annotateurs. Cela s'explique par la différence (statistiquement significative) de densité d'annotations (nombre d'annotations par token) : 0,16 pour les minutes et 0,08 pour les transcriptions. En effet, les commentateurs sportifs ne parlent pas uniquement des événements du matchs et ont tendance à digresser. En revanche, si l'on rapporte le temps d'annotation au nombre d'annotations produites, aucune différence n'est constatée entre minutes et transcriptions. Les différences de temps entre les deux modalités s'expliquent donc uniquement par le nombre plus important d'annotations à produire à volume de texte constant.

3 Analyse statistique des annotations

3.1 Mesures d'accord et estimation des "annotables"

Les calculs d'accords inter- et intra-annotateur servent à quantifier la fiabilité, et donc la qualité, des annotations produites, mais aussi à fixer une limite supérieure aux performances que l'on peut attendre d'un système automatique, et enfin, dans notre cas, à mesurer la difficulté de la tâche selon la modalité d'origine. Pour ce faire, les Kappa (κ) de Cohen (Cohen, 1960) et de Carletta (Carletta, 1996) sont préférés aux mesures plus simples telles que la F-mesure car ils normalisent l'accord observé en fonction de l'accord attendu (ou dû au hasard). Carletta considère que l'annotation par hasard se traduit par une unique distribution valable pour les deux annotateurs, alors que Cohen considère que ces distributions dépendent de chaque annotateur (Artstein et Poesio, 2008, pour une description complète et des comparaisons).

Cependant, ces définitions posent problème dès lors que ce ne sont pas seulement les étiquettes qui peuvent varier, mais aussi les éléments à annoter (les *marquables* ou *annotables*), puisqu'elles ne précisent en rien comment le désaccord sur les annotables doit être traité. Nous proposons donc d'étendre les κ en décomposant l'accord en un accord sur l'annotable et un accord sur l'étiquette. De telles mesures nécessitent donc de connaître le nombre d'*annotables* $\mathscr{M}$. Ce nombre d'annotables est évident ou connu *a priori* pour certaines tâches (comme l'étiquetage morphosyntaxique : tous les tokens sont annotables), mais ne peut être qu'estimé *a posteriori* pour des tâches comme la nôtre (Grouin *et al.*, 2011). Nous proposons pour ce faire une estimation originale basée sur une procédure EM (*Expectation-Maximization*) décrite dans l'algorithme 1. Celui-ci énumère itérativement le nombre d'annotables δ (étape de *Maximization*) en utilisant la probabilité γ (estimée itérativement) que tous les annotateurs aient manqué le même annotable, elle-même calculée grâce à l'estimation du nombre d'annotables δ de l'itération précédente (*expectation*).

Avoir une estimation la plus exacte possible du nombre d'annotables est un enjeu d'importance pour obtenir des accords inter-annotateurs réalistes. Par exemple, si l'on considère que tous

Algorithme 1 *Estimation EM des annotables*

Entrées : $\{\mathscr{M}_j\}$ (ensembles des éléments annotés par les annotateurs A_j) ; $\delta_0 = \left|\bigcup_j \mathscr{M}_j\right|$

for (i=1 ; $\delta_i \neq \delta_{i-1}$; i++) **do**

expectation : $\gamma_i = \prod_j P(A_j \text{ manque un marquable}) = \prod_j \frac{\delta_{i-1} - |\mathscr{M}_j|}{\delta_{i-1}}$

maximization : $\delta_i = \frac{\delta_0}{1-\gamma_i}$

end for

return δ

les mots (tokens) des textes sont des annotables (et donc ceux non annotés sont considérés annotés par défaut par une étiquette *sans-annotation*), le Kappa de Cohen pour les accords intra- et inter-annotateurs atteindrait respectivement 0,9456 et 0,9404, principalement par l'abondance des accords sur les très nombreux mots *sans-annotation*. De telles valeurs masquent des différences qui sont révélées avec l'estimation plus réaliste des annotables que nous proposons (voir sous-section 3.2).

Les deux κ, tels que nous les avons implémentés, sont aussi très stricts, puisque la moindre différence dans les annotations (étiquette bien sûr, mais aussi délimitation des entités) est considérée comme un désaccord. Quand cela est possible, nous fournissons donc également la mesure d'accord entropique implémentée dans GLOZZ (Mathet et Widlöcher, 2011) ; celle-ci autorise en effet les correspondances partielles d'annotation et fournit donc des valeurs d'accord prenant en compte ces accords partiels. Elle ne s'applique cependant pas encore aux relations.

3.2 Accords inter-annotateurs

Le tableau 3 présente l'accord inter- et intra-annotateur, selon la modalité, calculés avec le κ de Cohen, et, à des fins de comparaison, la mesure d'entropie de GLOZZ. Le κ de Carletta a également été calculé et est très proche dans la quasi-totalité des cas au κ de Cohen ; nous ne reportons donc pas ses valeurs par manque de place. Cette proximité signifie qu'il n'y a pas de biais d'annotateur : les distributions des annotations produites par chacun des annotateurs sont très similaires (Artstein et Poesio, 2008). On constate sans surprise que l'accord (aussi bien inter- qu'intra-annotateur) a tendance à être plus faible dans les transcriptions que dans les minutes, à l'exception d'une transcription pour laquelle les unités/actions ont produit un accord bien supérieur (près de 0,65). Cette tendance générale se manifeste spécialement dans les cas d'annotations complexes comme les relations. Les spécificités de l'oral mentionnées précédemment, et en particulier le style elliptique propre aux commentaires, expliquent facilement cette différence.

Si le calcul d'accord inter-annotateurs est devenu une bonne pratique standard du développement de ressources annotées, nous souhaitons promouvoir dans cet article l'intérêt d'une analyse plus détaillée. Cela est d'autant plus important quand les éléments annotés relèvent de catégories différentes et que ces catégories elles-mêmes ont des populations très différentes, comme c'est le cas ici. En effet, les valeurs présentées précédemment masquent des disparités importantes entre catégories d'annotation. Dans le tableau 4, colonnes 2 et 5, nous développons les résultats d'accord inter-annotateurs par regroupements de catégories. Les difficultés accrues sur les transcriptions se vérifient à cette échelle, mais l'on constate en outre de très faibles accords pour

	inter-annotateurs		intra-annotateur A1		intra-annotateur A2	
	κ de Cohen	Glozz	κ de Cohen	Glozz	κ de Cohen	Glozz
Minutes unités/actions	0,5992	0,7627	0,7531	0,8753	0,7109	0,8519
Minutes relations	0,5707	-	0,6377	-	0,5983	-
Transcriptions unités/actions	0,6234	0,7498	0,7558	0,8327	0,6812	0,8179
Transcriptions relations	0,4345	-	0,4010	-	0,4701	-

TABLE 3 – Accords inter-annotateurs et intra-annotateur par modalité

	Minutes			Transcriptions		
	κ	Incertitude	Gain d'entropie	κ	Incertitude	Gain d'entropie
Acteurs	0,9228	0,5	-	0,8974	1,0	-1
Circonstants	0,4827	1,9	49	0,4441	10,0	15
Actions arbitrales	0,5999	4,3	-	0,5082	19,7	7
Actions autres	0,3240	1,3	92	0,1407	9,8	26
Relations arbitrales	0,6355	10,7	-	0,4520	18,4	8
Relations autres	0,5540	10,2	8	0,3793	69,9	23

TABLE 4 – Accords inter-annotateurs par modalité et par famille d'annotations

certaines catégories. Les accords sur les entités offrent un grand contraste entre les annotations des acteurs et des circonstants, davantage sujets à interprétations. De la même manière, les événements (actions ou relations) sanctionnés par une action de l'arbitre obtiennent des accords bien supérieurs aux autres événements. Un examen détaillé des résultats montre que les annotateurs sont rarement en désaccord sur les types des éléments annotés, mais qu'ils annotent des éléments différents. Ce dernier point justifie d'autant plus l'emploi de notre technique d'estimation des annotables et explique pourquoi la définition standard des κ sur-estime tant l'accord.

3.3 Incertitudes

Les annotateurs avaient la possibilité d'indiquer les annotations leur posant problème, pour quelque raison que ce soit, à l'aide d'un champ *Incertitude*. Ces incertitudes permettent, lors de la campagne, de préciser les instructions d'annotations, de comprendre certaines annotations lors de l'utilisation du corpus, mais aussi d'aider à l'analyse automatique des résultats, comme indicateur de la difficulté d'annotation. Il est à noter qu'un seul des annotateurs de la campagne a véritablement utilisé les incertitudes, mais de manière systématique.

Dans les colonnes 3 et 6 du tableau 4, nous présentons les taux d'incertitude par catégorie d'annotations et par modalité. On y constate encore une fois que proportionnellement plus d'incertitude concerne l'oral retranscrit (différence statistiquement significative, test de Student pour deux ensembles, avec $p = 0,05$).

Nous nous sommes intéressés au lien éventuel entre incertitude et désaccord. Nous avons cherché à savoir si la présence d'une incertitude est liée au désaccord. Par contre, nous considérons non interprétable l'absence d'incertitude. Pour ce faire, nous avons calculé la différence entre l'entropie de l'accord $H(Acc)$ (eqn 2) de la variable aléatoire Acc indiquant s'il y a accord ou non ($\mathscr{D}_{Acc} = \{vrai\,;\,faux\}$) et l'entropie conditionnelle de l'accord sachant qu'une incertitude est

présente ($H(Acc|Inc = présent)$, eqn 2). Un gain positif signifie que l'incertitude aide à discerner les accords des désaccords. Autrement dit, pour une catégorie donnée, un gain positif indique que l'incertitude peut aider à prédire les catégories susceptibles de désaccord.

$$H(Acc) = -\sum_{v \in \mathcal{D}_{Acc}} P(Acc = v) \log P(Acc = v) \quad (1)$$

$$H(Acc|Inc = vrai) = -\sum_{v \in \mathcal{D}_{Acc}} P(Acc = v|Inc = vrai) \log P(Acc = v|Inc = vrai) \quad (2)$$

Ces gains sont indiqués en colonnes 4 et 7 du tableau 4 pour les familles d'annotation (dans trois cas, il n'y a pas assez d'incertitudes pour les calculer). À une exception près, ils sont tous positifs, ce qui signifie que ces incertitudes sont des bons indicateurs d'erreurs, même si elles n'ont été posées que par un seul annotateur. Que ce soit pour les minutes ou les transcriptions, il faut remarquer que le gain est d'autant plus fort que le taux de désaccord est important. L'étude des causes de ces incertitudes est donc une piste privilégiée pour la correction systématisée des désaccords et donc des éventuelles erreurs d'annotation.

4 Conclusion et perspectives

L'analyse de la campagne d'annotation présentée dans cet article a mis en exergue différents éléments. D'un point de vue méthodologique, notre technique d'estimation des annotables doit permettre un calcul d'accord inter-annotateurs plus réaliste dans les cas où leur nombre peut varier selon l'annotateur. Nous avons aussi montré que les bonnes pratiques ne sauraient se limiter à un calcul d'accord inter-annotateurs unique pour l'ensemble des annotations quand celles-ci relèvent de catégories différentes et d'effectifs non équilibrés. Enfin, nous avons montré que l'étude statistique des incertitudes met au jour une possibilité de détecter systématiquement les désaccords ou erreurs potentiels. Ces différentes analyses nous ont aussi permis de montrer que le coût d'annotation des textes issus de l'oral est moindre que pour ceux issus de l'écrit, du fait de la différence de densité des annotations. En revanche, les indicateurs de qualité (désaccord, incertitudes) indiquent sans ambiguïté la difficulté accrue de traiter de l'oral. Les annotations seront librement disponibles [3] dès que les corrections identifiées auront été effectuées. Le guide d'annotation mis à jour sera lui-aussi fourni.

En suite de ce travail, et aussi bien d'un point de vue théorique que pratique, nous souhaitons développer des approches permettant de propager automatiquement des corrections d'annotations à partir de quelques corrections apportées à une petite quantité de données. Ces approches s'appuieraient d'une part sur les analyses précédentes pour détecter les catégories les plus problématiques, et éventuellement sur des approches d'apprentissage artificiel pour proposer des corrections.

Références

ARTSTEIN, R. et POESIO, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

3. sous licence LGPL-LR à `http://www.irisa.fr/texmex/people/claveau/corpora/FootQuaero/`.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. *In Actes de First International Conference on Language Resources and Evaluation (LREC 1998)*, Grenade, Espagne.

BONNEAU-MAYNARD, H., ROSSET, S., AYACHE, C., KUHN, A. et MOSTEFA, D. (2005). Semantic annotation of the french media dialog corpus. *In Actes de InterSpeech*, Lisbonne, Portugal.

CARLETTA, J. (1996). Assessing Agreement on Classification Tasks : the Kappa Statistic. *Computational Linguistics*, 22:249–254.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

DANDAPAT, S., BISWAS, P., CHOUDHURY, M. et BALI, K. (2009). Complex Linguistic Annotation - No Easy Way Out ! A Case from Bangla and Hindi POS Labeling Tasks. *In Proceedings of the third ACL Linguistic Annotation Workshop*, Singapour.

FORT, K. et CLAVEAU, V. (2012). Annotating football matches : Influence of the source medium on manual annotation. *In Actes de Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie. 6 pages.

FORT, K., NAZARENKO, A. et RIS, C. (2011). Corpus linguistics for the annotation manager. *In Actes de Corpus Linguistics*, Birmingham, Angleterre. 13 pages.

GASIGLIA, N. (2003). Pistes méthodologiques pour l'exploration d'un corpus à haut rendement relatif au parler du football, une langue de spécialité de grande diffusion. *In 3es journées de linguistique de corpus*. Centre de Recherche en Littérature, Linguistique et Civilisation (CRELLIC), Université de Bretagne-Sud, Lorient.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K., GALIBERT, O. et QUINTARD, L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. *In Actes de 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

GUT, U. et BAYERL, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. *In Actes de Speech Prosody*, pages 565–568, Nara, Japon.

MATHET, Y. et WIDLÖCHER, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. *In Actes de Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, Montpellier, France.

NEMRAVA, J., SVATEK, V., SIMUNEK, M. et BUITELAAR, P. (2007). Mining over : Football match data : Seeking associations among explicit and implicit events. *In Proc. of Znalosti 2007*.

SCHMIDT, T. (2008). *The Linguistics of Football (Language in Performance 38)*, volume 38, chapitre The Kicktionary : Combining corpus linguistics and lexical semantics for a multilingual football dictionary, pages 11–23. Gunter Narr, Tübingen, Allemagne.

VOORMANN, H. et GUT, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

WIDLÖCHER, A. et MATHET, Y. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *In Actes de Traitement Automatique des Langues 2009 (TALN 2009)*, Senlis, France.