Annotating Football Matches: Influence of the Source Medium on Manual Annotation

Karën Fort*, Vincent Claveau†

*INIST - CNRS / LIPN
2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy, France
karen.fort@inist.fr

† IRISA - CNRS
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr

Abstract

In this paper, we present an annotation campaign of football (soccer) matches, from a heterogeneous text corpus of both match minutes and video commentary transcripts, in French. The data, annotations and evaluation process are detailed, and the quality of the annotated corpus is discussed. In particular, we propose a new technique to better estimate the annotator agreement when few elements of a text are to be annotated. Based on that, we show how the source medium influenced the process and the quality.

Keywords: Manual corpus annotation, Inter-annotator agreement, Football, Speech Transcription

1. Introduction

It has become a cliché to state that multi-modal and multimedia documents are now widely spread and part of our every day life. Yet, proposing intelligent processing of these documents or simply accessing the information they contain is still an issue for real-world applications. Moreover, building and annotating resources to develop and test such applications raises issues that are seldom documented.

In this paper, we present the development of an annotated text corpus of football[1] matches built from video (speech transcripts) and specialized websites. In that respect, the corpus is not multi-modal *per se*, as only text is processed, but it contains data from both written and oral sources. This corpus and the associated annotations were created in order to develop and test automatic tools for video summarizing, re-purposing or event extraction from football broadcasts. This application, developed with an industrial partner, is not further developed in this article, but it has an important impact on the elements to be annotated (see Section 3. and Fort et al. (2009)).

In this paper, in addition to introduce a new resource, we aim at two other goals. First, we detail the annotation campaign and how good practice rules were implemented for this heterogeneous corpus. Secondly, across the annotation process, we exhibit quantitative and qualitative differences in results between the written and oral sources. To do so, we use inter-annotator and intra-annotator agreement measures and adapt existing ones to 1) acknowledge the quality of the produced annotation, 2) outline the differences in the quality results between the written and oral (from video) sources.

The article is structured as follows. After a brief review of related studies in Section 2., we present the corpus, its annotation, and the campaign in Section 3. We then detail agreement measures and their results, used for the resource development, in Section 4., before concluding.

[1] in the sense of soccer.

2. Related Work

A number of publications dealing with football-based applications (Nemrava et al., 2007, for example) refer to a domain annotated corpus. However, to our knowledge, none of them describe in details the manual annotation of the corpus itself. Besides, none of them concern a French corpus. Other studies used football corpora to create more or less detailed monolingual (Nathalie Gasiglia, 2003) or multilingual (Schmidt, 2008) lexicons. In these cases, if the associated publications detail the annotation of the corpora that were used, the annotation itself was more linguistically-oriented than domain-oriented and therefore raised different issues.

3. Campaign Preparation

3.1. Data Preparation

The corpus we used covers 16 European football matches. It is made of 24 transcripts of the video commentaries of the matches (1 per half-time, for 12 matches) and 16 files containing the written minutes of matches (the same 12 matches that are covered by the transcripts and 4 additional matches). The total size of the corpus reaches about 250,000 tokens. As shown by Fort et al. (2011), its main characteristic is to be very heterogeneous, be it from the point of view of the type of matches (French 1^{st} league matches, international matches, etc), of the files size (from 1,116 tokens per match for minutes to 21,000 tokens per match for the transcripts), or of the sources (different TV channels and commentators).

The speech contained in the video was manually transcribed using TRANSCRIBER (Barras et al., 1998) and its default guidelines. It is worth noting that the transcripts are aligned with the speech. Therefore, it provides us with a precise timestamp for each word and annotation in the transcripts. The minutes also contain time information since

each action is preceded by its occurring time in the match. Thus, it makes it possible to map events described in both sources.

3.2. Annotations

We decided to decompose the annotation into three steps, corresponding to layers of analysis of growing complexity, that were easy to annotate simultaneously. Thus, the annotators had to first annotate all the units (named entities, time and location), then the actions and finally the relations (see details in table 1[2]).

These labels were selected in three steps: 1) selection from an available, and rather general, football ontology (Crampes and Ranwez, 2000) from Ecole des Mines d'Alès[3], keeping our application in mind as recommended in (Leech, 2005), 2) modifications following the training phase, 3) modifications following the pre-campaign.

As the corpus is made of two different media, one of which, the transcripts, is ellipses-prone ("`Makoun. Et c'est récupéré. Clerc, avec Cris. Boumsong, Makoun.`" [*Makoun. Saved. Clerc with Cris. Boumsong, Makoun.*]), we decided not to annotate the actions' predicate, but the actors. The same goes for the relations. This choice was made in order to maintain only one annotation guide and a homogeneous annotation process. However, and this is especially the case in transcripts, the actor of an action does not always appear in the text: "`Grand dribble en pivot bien pris`" [*Large dribble in pivot well stopped*] (here, the actor of the dribble is not indicated). The same goes for the relations, in which the source or/and the target actor can be omitted: "`Ribéry avec une faute sur Gourcuff`" [*Ribery with a foul on Gourcuff*] (here, the source actor of the foul is missing). In these cases, we asked the annotators to anchor the annotation on the action predicate and to add a predefined feature (*Missing Actor* for actions and *Missing source/target* for relations). Note that for the relations, we had to add a new unit, *ActionPourActeurVide* (ActionForEmptyActor), used to annotate the predicate, in order to anchor the relation on it.

3.3. Methodology

The annotation was performed by two annotators from INIST-CNRS, one man and one woman, both experts in football (regular player and former player). We chose to use GLOZZ (Widlöcher and Mathet, 2009) as annotation tool: it is easy to use and supports the annotation of relations. The files to be annotated were dispatched between the annotators so that they had a similar workload, taking into account the types of files (league 1 matches, international matches, etc), their source (minutes or transcripts) and their size. Apart from the training part, the corpus was automatically pre-annotated for player and coach names, using lists found on specialized websites. Part of the annotation work was therefore about correcting pre-annotations, which proved to help gaining time and quality in at least the annotation of part-of-speech (Fort and Sagot, 2010). We also advised the annotators to work first on the match minutes before annotating the transcripts (supposedly more ambiguous), when available. It is important to note that we finally decided to annotate the transcripts directly, not using the video, in order to gain time (more than 2 hours per transcript).

Annotators were asked to annotate layer by layer (see section 3.2.), and to track their time for each file and each annotation step within the file, using a freely available on line tool[4]. They were also asked to work at least 10 hours a week on the annotation and keep a steady rhythm at it to optimize the learning curve and the quality of the work, but this recommendation was not always followed, due to busy schedules. We also told the annotators not to hesitate to add comments and we added an *Uncertainty* feature to the annotations that they could use. Annotator 1 used both possibilities while Annotator 2 did not.

We used the methodology described by Bonneau-Maynard et al. (2005) and computed the inter-annotator agreement early in the process in order to check for inconsistencies in the annotation model and obvious ambiguities in the tagset to improve the annotation guidelines. We also computed intra-annotator agreements, as recommended by Gut and Bayerl (2004).

The annotation campaign itself was done in several phases: 1) training: on the smallest match minutes file (not pre-annotated), using the annotation tool, 2) pre-campaign I: annotation by both annotators of the same corpus sample (of match minutes), computation of the inter-annotator agreement, discussion about disagreements, update of the guidelines, 3) pre-campaign II: annotation by both annotators (together) of one minutes file, new update of the guidelines, 4) pre-campaign III: annotation by both annotators of the same corpus sample (of transcripts), computation of the inter-annotator agreement, discussion about disagreements, new update of the guidelines, 5) campaign: annotation by annotators of the files assigned to them (match by match). Finally, other inter-annotator agreements and intra-annotator agreements were computed at the end of the campaign.

4. Results and analysis

4.1. Agreement measures

Computing intra- and inter-annotator agreements is essential when developing annotated resources: it is used to assess the reliability, hence the quality of the produced annotations, to set an upper bound of the performance of automatic systems, and in our case, to highlight the difficulty of the task according to the source modality. Cohen's (Cohen, 1960) or Carletta's (Carletta, 1996) κ are preferred to simpler measure like F-measure since they take into account the chance agreement (Artstein and Poesio, 2008, for a complete description and comparison). Yet, such measures

[2]Note that the grouping of categories into *actors* and *circumstants*, then *initiated by referee* and *others* presented in table 1 was only defined *a posteriori* for the evaluation and did not exist in the data model used by the annotators.

[3]`http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml`

[4]TIMETRACKER (`http://www.formassembly.com/time-tracker/#`).

Units	
actors	*Joueur* (Player), *Equipe* (Team), *Arbitre* (Referee), *Entraineur* (Coach), *ArbitreAssistant* (AssistantReferee), *Président* (President)
circumstants	*EspaceSurTerrain* (LocationOnField), *LieuDuMatch* (MatchLocation), *TempsDansMatch* (TimeInMatch)
Actions	
initiated by referee	*TirerCoupFrancDirect* (DirectFreeKick), *TirerCoupFrancIndirect* (IndirectFreeKick), *TirerCorner* (Corner), *TirerPenalty* (Penalty), *FaireFauteDeJeu* (Foul), *HorsJeu* (Offside), *MarquerBut* (ScoreGoal), *PrendreCartonJaune* (YellowCard), *PrendreCartonRouge* (RedCard), *PrendreRappelALOrdre* (Warning)
others	*Centrer* (Center), *FaireTentative2Centre* (CenterAttempt), *Dribbler* (Dribble), *RaterBut* (MissGoal), *ArreterBut* (StopGoal), *IntercepterBallon* (Interception), *PossederBallon* (HaveBall), *ActionDuPublic* (AudienceAction)
Relations	
initiated by referee	*FaireFauteSurJoueur* (FoulOnPlayer), *TaclerFaute* (FoulTackle), *RemplacerJoueur* (ReplacePlayer)
others	*FaireCombinaison* (Combination), *FairePasse* (Pass), *FaireTentative2Passe* (PassAttempt)

Table 1: Annotation steps and corresponding labels

require to evaluate the number of *markables* (entities that may require to be annotated). While the number of markables is obvious and known *a priori* for some tasks (like PoS tagging), it can only be estimated *a posteriori* for annotation tasks like ours (Grouin et al., 2011). To overcome this issue, we propose an *a posteriori* estimation based on the expectation-maximization procedure described in Algorithm 1. It iteratively estimates the number of markables δ (maximization step) using the (iteratively estimated) probability γ that all the annotators miss a same markable computed as the product of probability of A_j missing a markable (expectation step). In the following subsection, we use this algorithm to estimate the number of markables when computing Cohen's and Carletta's κ.

Algorithm 1 *EM Algorithm*

Input: $\{M_j\}$ (sets of marked elements by annotators A_j)

$$\delta_0 = \left| \bigcup_j M_j \right|$$

for (i=1 ; change in δ ; i++) **do**

expectation: $\gamma_i = \prod_j P(A_j \text{ misses a markable}) = \prod_j \frac{\delta_{i-1} - |M_j|}{\delta_{i-1}}$

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

end for

return δ

For instance, the intra- and inter-annotator Cohen's κ for the annotation of entities and actions in the minutes, when computed (as usual) by considering all the tokens as markables, respectively reaches 0.9456 and 0.9404. Such high values mask differences that are better emphasized when computing κ with the estimation of markables that we propose (see Sub-Section 4.4.). The two κ, as implemented, are also very strict in the sense that any slight difference between two annotations will be considered as a disagreement. Thus, when possible, we also provide the entropy agreement values as defined by Mathet and Widlöcher (2011) and implemented in GLOZZ. This measure takes into account partial matches and thus provides less pessimistic agreement values, but it does not apply to relations.

4.2. Data on Process

The choice of the annotation tool has an important impact on the annotation campaign. Our data model was designed to comply with GLOZZ's constraints. Thus, we used no relation with more than two actors and marked the actions with a prefix ("A_") to distinguish them from simple units. Also, as GLOZZ does not allow for the direct modification of the source text, the annotators could not correct the typographic errors, the missing whitespaces or the tokenization problems that occurred in the corpus, in particular in the transcripts. Annotator 1, who inserted a lot of comments, noted 94 of them, all in the transcripts (vs. 1 for annotator 2, in a minutes file). These transcription errors also impacted the automatic pre-annotation: annotator 1 noticed that 321 named entities were not pre-annotated due, in particular, to typographic errors. However, we obtain an inter-annotator agreement (using GLOZZ Entropy measure) between the annotators and the automated pre-annotation of more than 0.9 on transcripts and 0.8 on minutes.

The total number of annotations added or corrected by the annotators is 37,784, 27,736 of which (i.e. more than 73%) were added or corrected in the transcripts. All the categories were used, but two of them only twice (*TirerCoupFrancIndirect* and *TirerPenalty*) and only in the minutes, and the annotators found only 6 red cards (*PrendreCartonRouge*) and 9 *President*.

As for the missing actors, the annotators found 586 of them in the actions and 404 in the relations (190 source actors, 173 target actors and 41 source and target actors). The majority of these missing actors appear in the transcripts (304 out of 586). This is consistent with the comments made by the annotators in which they note a lot of doubts on the *FairePasse* relation (nearly 800) and, more generally, on

what is going on in the transcripts (they noted 1,429 uncertainties in the transcripts files out of a total of 1,505).

4.3. Annotation time

Table 2 presents the mean annotation times (per 1,000 tokens) for each annotator and source. In order to check if the differences are significant, we ran statistical tests (Welsh two sample t-test, with p=0.05). These tests show that there are no statistically significant differences between the annotation time of the annotators, both for minutes and transcripts. More interestingly, the differences between the modalities are proved statistically significant, for both annotators, when considering the time spent by token. But we also find that no statistically significant difference is found between the annotation time by annotation produced between the written and the oral modality. These two significance results may seem contradictory, but it is simply explained by the (statistically significant) difference of density of annotations (number of annotations given the number of tokens); the mean density for minutes is 0.16 while those of transcripts is 0.08. Indeed, video commentators tend to make small talk or talk about other events during the match, thus diluting interesting information.

	Minutes	Transcripts
Annotator 1	36.92	20.03
Annotator 2	41.30	16.06

Table 2: Mean annotation time by source and annotator, in minutes/1,000 tokens

4.4. Annotation agreements

Table 3 presents the intra- and inter-annotator agreement values with Cohen's κ, Carletta's κ and the GLOZZ Entropy measure for the different layers of annotation. Several figures are noteworthy. First of all, Cohen's κ and Carletta's κ are very close in almost all cases. It means that there is no annotator bias, i.e. the distributions of annotations produced by the annotators are very similar (Artstein and Poesio, 2008). Secondly, the 3 different measures show that annotating relations is more error-prone than annotating unary annotations (units and actions).

Computing annotation agreements has become a standard when developing annotated resources, but in this paper, we would like to promote the interest of a finer grain analysis. This is especially important when the elements to annotate belong to different categories, and when these categories comprise very different population, as in our case. Indeed, more detailed results presented in table 4 show that significant disparities between the annotation categories actually exist. They also highlight the need for post-processing for certain categories of actions.

Interestingly, the absence of bias between the annotators is also verified at this level, as well as the higher difficulty of processing transcripts. If some categories yield very low agreement measures (eg. *PosséderBallon* (HaveBall), *FaireTentative2Passe* (PassAttempt)), events (actions or relations) initiated or validated by an action of the referee are less open to interpretation and thus obtain better results than other events. Similarly, agreement on entities show a high contrast between the *actors* and the *circumstants*.

From a qualitative point of view, a closer analysis of the disagreements shows that the annotators rarely disagree on the type of an annotation, but annotate different elements. Last, unsurprisingly, the agreement values (both inter- and intra-annotator) tend to be lower in transcripts than in minutes. It is especially the case with complex annotations like relations. The previously mentioned oral specifics, in particular ellipses, easily explain this difference.

4.5. Qualitative analysis

Based on the quantitative results presented in the previous subsections, the principal causes of disagreements were searched for the most error-prone annotation categories, in one file of minutes and two files of transcripts. This analysis was tedious but very useful; it made it possible to build the following typology of the main causes of disagreements:

- errors due to the a misuse of the annotation tool (eg. units annotated 2 times);
- over-annotation or under-annotation of an annotator;
- disagreements on the frontiers of the annotated linguistic unit;
- disagreements on the anchoring of a relation;
- ambiguities, especially in speech transcripts.

In the first three cases, the disagreements are due to an error of one of the annotators. The errors caused by a misuse of the tool are not frequent. On the contrary, the over-annotation of some linguistic phenomena is more frequent, but can be controlled by adding recommendation in the annotation guidelines. The under-annotation and forgotten annotations are more difficult to detect and to solve, since they are mainly caused by lapse of concentration.

The last two types of errors are more complex to handle, as they are not errors *per se*. For instance, concerning the relation anchoring, both annotators often identified the right actor, but not the same occurrence of its name (although the guidelines gave directions to prevent this). This is what happened in the example presented in Figure 1.

Last, ambiguities, mainly found in speech, made the annotation of the transcripts tedious, implying to re-read several times the same sentence. Despite those efforts, many doubts on the annotation may still persist. For instance, one could think that the *MarquerBut* (ScoreGoal) action, which is very important from an applicative point of view, is fairly easy to annotate, but in the example given in Figure 2, the speech ambiguities misled the annotator to indicate that `Gouffran` scored, while in fact `Gourcuff` scored.

The results of this detailed analysis suggest different ways to improve the quality of this annotated corpus. First, the fusion of the annotations from the two annotators, possibly corrected by one of them, would provide a more complete and stable result. Additionally, the intra-annotator agreement shows that annotator 2 was less coherent with himself than annotator 1 (except on relations, in the transcripts). This justifies that his/her annotations be reviewed and possibly corrected in priority. Additionally, annotator 1 could

Inter-annotator agreement			
	Cohen's κ	Carletta's κ	GLOZZ
Minutes units/actions	0.5992	0.5991	0.7627
Minutes relations	0.5707	0.5707	-
Transcripts units/actions	0.5979	0.5879	0.7645
Transcripts relations	0.4050	0.4025	-
Transcripts units/actions	0.6490	0.6490	0.7351
Transcripts relations	0.4640	0.4635	-
Intra-annotator agreement			
	Cohen's κ	Carletta's κ	GLOZZ
Minutes units/actions A1	0.7531	0.7531	0.8753
Minutes relations A1	0.6377	0.6377	-
Minutes units/actions A2	0.7109	0.7109	0.8519
Minutes relations A2	0.5985	0.5983	-
Transcripts units/actions A1	0.7558	0.7558	0.8327
Transcripts relations	0.4010	0.3904	-
Transcripts units/actions A2	0.6812	0.6812	0.8179
Transcripts relations	0.4701	0.4700	-

Table 3: Annotation agreement values

	Minutes		Transcriptions	
	Cohen's κ	Carletta's κ	Cohen's κ	Carletta's κ
Actors	0.9228	0.9228	0.8974	0.8973
Circumstants	0.4827	0.4826	0.4441	0.4440
Actions init. by referee	0.5999	0.5999	0.5082	0.5082
Other actions	0.3240	0.3240	0.1407	0.1403
Relations init. by referee	0.6355	0.6354	0.4520	0.4503
Other relations	0.5540	0.5540	0.3793	0.3789

Table 4: Annotation agreements by modality and annotation category

be used as a corrector, after sufficient training with the updated guidelines.

5. Conclusion

This article presents in details the manual annotation process and quality of a football match annotation campaign. The produced annotations are freely available[5] as well as the annotation guidelines, in French. At the heart of this annotation process is the evaluation of annotator agreement. We proposed a new and simple way to estimate the number of markables, which is a key element in usual annotator agreement measures like Cohen's κ, and ensures not to obtain over-optimistic results. Different perspectives are foreseen for this work.

First, the qualitative analysis of the corpus is still ongoing and will probably lead to another version of the annotations, with corrections. Secondly, from a more multi-modal point of view, the differences of results between the oral and written sources will also be investigated, and should lead to interesting insights both from a linguistic and applicative points of view.

[5]under LGPL-LR license at `http://www.irisa.fr/texmex/people/claveau/corpora/FootQuaero/`.

6. Acknowledgments

We want to thank the annotators of the campaign, Claire Ris and Alain Zérouki from INIST-CNRS, for their hard work and their rich feedback, as well as Ali-Reza Ebadat, INRIA-INSA, Véronika Lux-Pogodalla, ATILF-CNRS, and Claire-Hélène Demarty, Technicolor Rennes, for helping us to collect and process the data. This work was realized as part of the Quæro Programme[6], funded by OSEO, French State agency for innovation.

7. References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998)*, Las Palmas, Spain, May.

Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic Annotation of the French Media Dialog Corpus. In *Proceedings of the InterSpeech Conference*, Lisboa, Portugal.

[6]`http://quaero.org/`

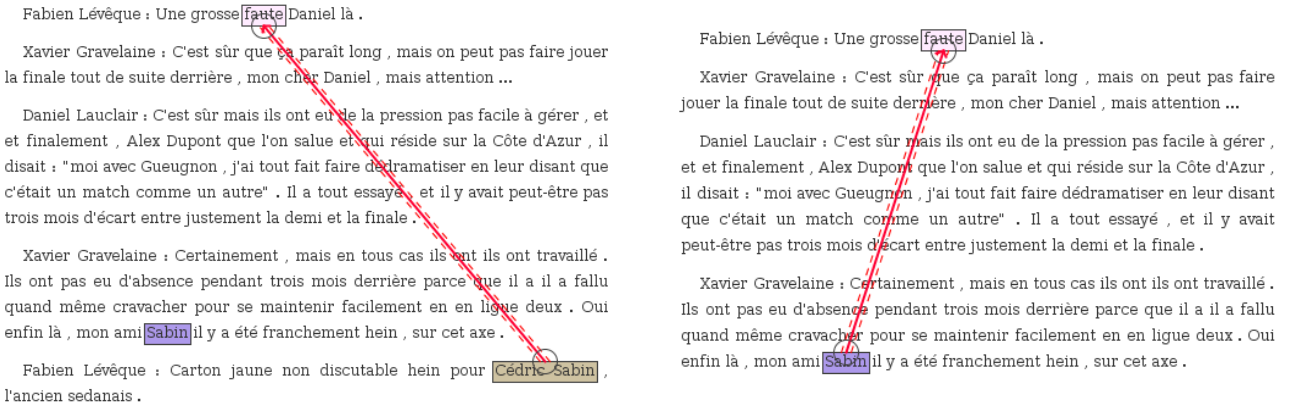

Figure 1: Example of disagreement on the anchoring of a relation (*FaireFauteSurJoueur*/*FoulOnPlayer*)

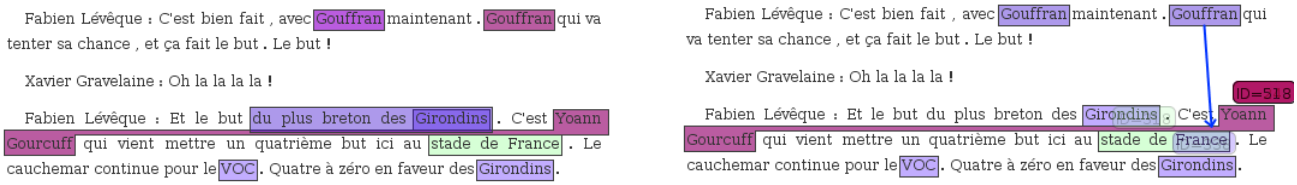

Figure 2: Example of disagreement due to speech ambiguities (*MarquerBut*/*ScoreGoal* action)

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, 22:249–254.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Michel Crampes and Sylvie Ranwez. 2000. Ontology-supported and ontology-driven conceptual navigation on the world wide web. In *Proceedings of the 11th ACM Conference on Hypertext (HT'00)*, San Antonio, Texas, USA.

Karën Fort and Benoît Sagot. 2010. Influence of Pre-annotation on POS-tagged Corpus Development. In *Proceedings of the 4th ACL Linguistic Annotation Workshop (LAW)*, Uppsala, Sweden.

Karën Fort, Maud Ehrmann, and Adeline Nazarenko. 2009. Towards a Methodology for Named Entities Annotation. In *Proceedings of the 3rd ACL Linguistic Annotation Workshop (LAW III)*, Singapore.

Karën Fort, Adeline Nazarenko, and Claire Ris. 2011. Corpus linguistics for the annotation manager. In *Proceedings of the Corpus Linguistics Conference*, Birmingham, England.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th ACL Linguistic Annotation Workshop (LAW)*, pages 92–100, Portland, Oregon, USA.

Ulrike Gut and Petra Saskia Bayerl. 2004. Measuring the Reliability of Manual Annotations of Speech Corpora. In *Proceedings of the Speech Prosody*, pages 565–568, Nara, Japan.

Geoffrey Leech, 2005. *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Adding Linguistic Annotation, pages 17–29. Oxford: Oxbow Books.

Yann Mathet and Antoine Widlöcher. 2011. Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In *Proceedings of the Traitement Automatique des Langues Naturelles 2011 (TALN 2011)*, Montpellier, France.

Nathalie Gasiglia. 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 *Proceedings of the 3es journées de linguistique de corpus*, Lorient, France.

Jan Nemrava, Vojtech Svatek, Milan Simunek, and Paul Buitelaar. 2007. Mining over: Football match data: Seeking associations among explicit and implicit events. In *Proceedings of the Znalosti 2007*.

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

Antoine Widlöcher and Yann Mathet. 2009. La plate-forme glozz : environnement d'annotation et d'exploration de corpus. In *Proceedings of the Traitement Automatique des Langues 2009 (TALN 2009)*, Senlis, France.