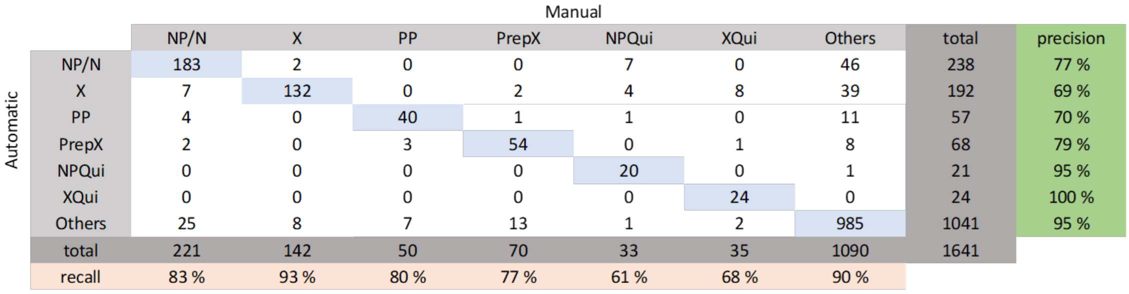
Syntactic Analysis of Live Sportscasts: Developing Less Time-consuming Strategies

Anna Kupść, Gilles Boyé & Catherine Mathon
(Université Bordeaux Montaigne and CLLE Montaigne)

Keywords: sports commentaries, syntactic structures, nominal structures, syntactic parsing, multilingual corpus

The aim of this paper is to parse syntactically a large multilingual corpus of live sports commentaries[1] (LSC) to study the correlations between syntactic, prosodic and lexical data with the sports events. We explore the possibilities to replace time consuming manual annotation by UDPipe[2], a multilingual syntactic dependency parser. The results presented rely on our only rugby match (in French) analysed manually. We focus on the nominal structures specific to LSC discussed in (Augendre et al., 2018) :

- NP/N: noun phrases with (NP) or without an article (N): 'mêlée à introduction française' *scrum with French introduction*
- X: proper names: 'Jauzion'
- PrepX/PP: prepositions followed by proper names (PrepX) or prepositional phrases (PP) 'avec Heymans' *with Heymans*
- XQui/NPQui: proper names or noun phrases followed by relative clauses 'Hernandez qui insiste' *Hernandez who insists* (XQui)



		Manual								
		NP/N	X	PP	PrepX	NPQui	XQui	Others	total	precision
Automatic	NP/N	183	2	0	0	7	0	46	238	77 %
	X	7	132	0	2	4	8	39	192	69 %
	PP	4	0	40	1	1	0	11	57	70 %
	PrepX	2	0	3	54	0	1	8	68	79 %
	NPQui	0	0	0	0	20	0	1	21	95 %
	XQui	0	0	0	0	0	24	0	24	100 %
	Others	25	8	7	13	1	2	985	1041	95 %
	total	221	142	50	70	33	35	1090	1641	
	recall	83 %	93 %	80 %	77 %	61 %	68 %	90 %		

The overall performance is below NLP standards but still very helpful to extend annotation to our whole corpus. The error analysis revealed two kind of issues:

- tagging:
    - some terms such as 'allez' (*go*.imperative.2pl), 'mêlée' (*scrum*) or 'lancer' (*throw*) which are tagged as verbs by UDPipe have special values in LSC: 'allez' is an encouragement interjection, 'mêlée' and 'lancer' are nouns. Pre-tagging such terms would avoid these errors on category.

[1] Our data consists of about 97 hours of recordings in French, English and Japanese.
[2] https://lindat.mff.cuni.cz/services/udpipe/

    - other problems arise from particularities of French (e.g., ‘du’ is either a partitive article or a contracted preposition, which provides two different structures); replacing the tagger or fine-tuning it may avoid these problems
- segmentation:
    - the segmentation into IPUs[3] can split syntactic structures into a leading structure and follow-ups (BIS-structures) as in (1) with NPQui and NPQuiBIS:

      1.
         a. ‘les anciens Pelous Ibañez’ NPQui
            the old ones: Pelous, Ibañez
         b. ‘qui étaient déjà là en quatre-vingt dix-neuf’ NPQuiBIS
            who were already there in 99

      The manual NPQui structure covers two IPUs: the antecedent (1a) and the relative clause itself (1b). For UDPipe, as is, (1a) is just a NP rather than NPQui and (1b) is not analysed as a relative clause because there is no antecedent. The mismatches between the structure of individual IPU and manual labels call for a macro-analysis of (1) which groups the main and follow-up IPUs including BIS-structures as chunks in the main structure.[4]
    - meanwhile some IPUs contain more than one structure but, for lack of punctuation, they are problematically analysed as a single structure. We could use Whisper (Radford *et al.,* 2022), an ASR, to provide punctuation in the spirit of (Deulofeu 2011).

References

Augendre, S., Kupść, A., Boyé, G. and Mathon, C. (2018), Live TV sports commentaries: specific syntactic structures and general constraints (2018), in Legallois, D., Charnois, T. & Larjavaara, M. (Dir.), The Grammar of Genres and Styles: From Discrete to Non-Discrete Units. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110595864

Deulofeu, H.-J. (2011). Peut-on établir un système de ponctuation des transcriptions de textes oraux linguistiquement fondé ? Les propositions du groupe Rhapsodie. Langue française, 172, 115-131. https://doi.org/10.3917/lf.172.0115

Radford, A., Kim, J.-W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2022), Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356v1 [eess.AS] 6 Dec 2022

[3] Inter-pausal units, segments separated by minimum 200ms pauses.

[4] BIS-structures, excluded by Augendre et al. (2018), represent 904 IPUs of the 2545 in the corpus.