Teaching Syntax with Clarin Corpora and Resources

Antonio Balvet

Teaching Syntax with Clarin Corpora and Resources

Antonio Balvet
UMR 8163 – STL "Savoirs Textes Langage" – F-59000
Lille University, France
antonio.balvet@univ-lille.fr

Abstract

The recent COVID-19 pandemic has brought online learning to the forefront for learners and teachers. As a consequence, the demand for self-paced and adaptive learning resources has reached unprecedented levels. Prior to the virus outbreak and consecutive lockdowns, universities had been using Moodle (and other SCORM[1] compliant platforms) as a Learning Management System (LMS), which has helped make the transition from on-site to online learning. But teachers still have had to face the challenge of designing and implementing assessment activities in the form of self-correcting activities (true/false, multiple answer questions, mark the words, fill in the blanks questions, etc.), instead of plain printed quizzes and tests. This step has proved to be a major hurdle since designing, and most of all, manually editing formative and evaluative assessment activities is a very labor-intensive task. In this article, we present a framework that builds upon corpora and resources available from the LINDAT / CLARIAH-CZ Data & Tools platform in order to generate quizzes and other activities related to syntax, for the Moodle platform. After some background on using Natural Language Processing (NLP) and electronic corpora for teaching syntax, we present our corpus-to-quiz processing chain, and we outline preliminary results on deploying automatically generated French syntax quizzes in the classroom.

1 Introduction

The recent COVID-19 pandemic has emphasized the necessity of self-paced and adaptive learning resources. Even though universities around the world had been using e-learning platforms prior to this event, teachers were still confronted with a very labor-intensive task, since designing and editing self-correcting assessment activities for potentially large groups of learners, in a distance-learning context, proved very time-consuming. Moreover, designing and implementing such assessment activities by hand is both error-prone and subjective, by nature.

In this article, we present a solution to optimize manual labor by relying on publically-available corpora. In the first section, we outline projects that have been using NLP solutions for teaching syntax. In the second section, we present our corpus-to-quiz processing chain, which ingests annotations present in corpora available from the LINDAT / CLARIAH-CZ Data & Tools platform, to generate syntax quizzes. Lastly, we report preliminary results on deploying such automatically generated quizzes, both for distance and on-site learning. Our presentation is centered on French, although the principle presented here is applicable to any Universal Dependencies CONLL-U formatted corpus, with minor adjustments.

1.1 Background: using parsers and annotated corpora for linguistic exercises and activities

Our corpus-to-quiz processing chain aims both at reducing manual edition to a minimum, and at overcoming the subjectivity (and errors) associated with manually-created exercises. Other projects have tried to address exactly those issues, in the past, such as (Bick, 2001, 2004; Uibo & Bick, 2005; Wijlff, 2006),

[1] Sharable Content Object Reference Model.

in the framework of the VISL Corpus project[2]. This project was based on a Categorial Grammar parser architecture, tailored to different languages. Based on this CG parser, large syntactically parsed corpora were set up, which allowed VISL consortium members to implement a platform very similar to the well known "Sketch Engine" (Kilgarriff et al., 2008). In addition to syntax-aware concordancers, members of the VISL consortium also devised an array of gamified exercises, based on the CG-parsed corpora, for different languages.[3]

More recently, other projects have integrated high-precision and robust parsers to automatically generate grammar exercises for French: (Colin, 2020; Perez-Beltrachini et al., 2012), in the framework of the LORIA-led METAL project.[4]

Our approach distinguishes itself from the aforementioned projects in that it builds upon manually-verified syntactic annotations, taken from reference corpora –such as the French Treebank or Sequoia– in order to generate quiz questions, which are ready to integrate into LMS[5] platforms such as Moodle[6]. Moreover, our approach targets undergraduate students, while the METAL project, for example, targets primary school pupils. Therefore, our approach focuses on the exercise generation aspect, for an audience of young adults; all authentication procedures and learning analytics logging are handled by the particular LMS being used.

2 A corpus to quiz processing chain

Our processing chain for generating self-correcting quizzes on French syntax relies on CONLL-U formatted corpora. At the time of writing, the French Treebank (FTB) (Abeillé et al., 2003) and the Sequoia corpus (Candito et al., 2014) are the only reference corpora, annotated following the Universal Dependencies guidelines (De Marneffe et al., 2021), available for French.[7] Our corpus-to-quiz processing chain is outlined in figure 1.

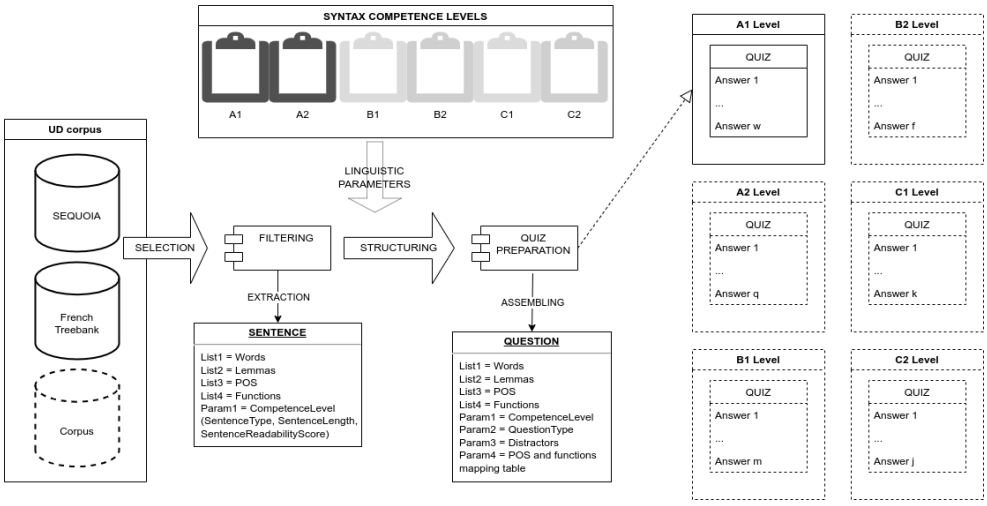

Figure 1: The corpus to quiz processing chain

Our tool is primarily targeted at university students attending introductory courses on syntax, as part of a curriculum in linguistics. We therefore need a definition of **syntax competence levels**,[8] which states

[2]"Visual Interactive Syntax Learning", Institute of Language and Communication (ISK), University of Southern Denmark (SDU) - Odense Campus: https://edu.visl.dk/visl2/

[3]For example, a syntactic labyrinth, as well as a "syntactic Tetris" and other syntactic games were implemented as (now obsolete) java applets.

[4]"Modèles et Traces au service de l'Apprentissage des Langues", Models and learning analytics for language learning.

[5]Learning Management System.

[6]The presented corpus to quiz processing chain is intended for Moodle, but other platforms could be targeted, as long as they allow importing quizzes and exercises from structured text files (XML, json, or other ad-hoc formats).

[7]Other CONLL-U formatted corpora are available for French, but they do not meet the same quality standards as the FTB and Sequoia. Moreover, they are mostly oral transcription corpora, which are not the best material for our purposes.

[8]This definition, inspired by the Common European Framework of Reference for Languages, is still a work in progress,

which syntactic features each learner profile is meant to acquire, from parts-of-speech, to constituent structure and functional relations. Based on the definition of syntax competence levels ranging from A1 (beginner) to C2 (advanced), a set of python scripts have been implemented for the automatic generation of Moodle quiz questions.[9] The scripts process sentences found in a set of CONLL-U corpora according to their overall syntactic types (e.g. simple vs complex syntactic structures, regular vs idiomatic units). Relevant sentences are then transformed into python data structures, and quiz questions are assembled by using each sentence's set of parameters, such as list of words, part-of-speech tag for each word, functions and dependency relations, etc. Different execution parameters allow the user to generate questions for different learner profiles. For example, the instructor can target specific words, lemmas and morphological constraints (e.g. a Noun bearing a *-able* suffix), specific subsets of part-of-speech, or function, tags, or even the number of distractors and syntactic annotation terminology to use (e.g. "SUJET" instead of "nsubj"). Figure 2 shows an example of a GIFT (General Import Format Template) structured quiz question on nouns bearing the *-able* suffix.

```
:: Parties du Discours ::[markdown]   Donner la partie du discours du mot **imputables** dans la phrase:
Très brièvement, il s'agit de limiter les émissions de CO2 **imputables** à l'homme.
{
        ~V_Subj
        ~PROPN
        ~PRO_Int
        ~Conj_de_Sub
        ~DET_Int
        ~V_Part_Prés
        ~Conj_de_Coord
        ~V_Inf
        ~AUX
        =ADJ
}
```

Figure 2: A quiz question on *-able* nouns

In this example, our question-generation script was launched with parameters to target all *-able* nouns. In the generated quiz question, learners must select the proper part-of-speech for noun "imputables" (*attributable*) in the context of the given sentence,[10] extracted from the Sequoia corpus. In this example, 9 distractors are shown (with a minimum of 2). The order in which distractors are presented, and other quiz parameters (e.g. randomized selection of individual questions, total allotted time) are defined by the instructor, for each quiz activity. In this prototype version, feedbacks must be manually provided by the instructor.[11]

Figure 3 shows how Moodle renders a GIFT-structured **part-of-speech quiz**, based on a sentence taken from the French Treebank.[12] Many French native speakers confuse the coordinating conjunction "ou" with the relative pronoun "où". Therefore, we have targeted "ou"/"où" and other typical confusing cases ("et"/"est", etc.) by stating a constraint on the form of the desired lemmas. These exercises are particularly adequate for the first weeks of a syntax course: the sentence is not too long, and the syntactic structure is relatively straightforward (a simple predicative structure). As such, it is adequate for less experienced learners (i.e. A1/A2 syntax competence level). Here, the quiz uses POS-tags available in the CONLL-U formatted version of the FTB, which are adapted and rendered so as to match a syntactic terminology closer to traditional grammar rather than UD categories. This mapping is achieved via a customizable equivalence table, it controls how the syntactic terminology will be presented to the learners, according to their competence level.[13]

In figure 4 a sentence[14] taken from the Sequoia corpus was used to assemble a quiz on **syntactic**

since no widely accepted, explicit definition of syntax competence levels could be found so far.

[9] All CONLL file preprocessing steps are performed thanks to the pyconll library. The code, as well as a large set of GIFT-structured questions are available at https://github.com/abalvet/ACE.

[10] *In a nutshell, the goal is to limit CO2 emissions attributable to mankind.*

[11] Chatbots, such as ChatGPT, might be integrated in future versions, in order to generate feedbacks based on learners's responses and competence levels.

[12] *In the plural form, since this event is economic, ecological, ideological and even iconoclastic.*

[13] Less advanced L1 students can be presented with a rather classical set of grammatical distinctions while more advanced L3 students can be exposed to the terminology and syntactic distinctions following actual UD guidelines.

[14] *On this matter, we ask ourselves the question of why feasibility studies and technical assistance measures amount to 47%*

Figure 3: A parts-of-speech quiz generated from a FTB sentence

functions, by leveraging on the dependency annotations available in the CONLL-U formatted version of the corpus. Here, the expected answer is "COD" (direct object), since *question* is the nominal head of the NP governed by *posons* (ask). As can be seen, the particular question shown is part of a quiz activity comprising 40 questions.

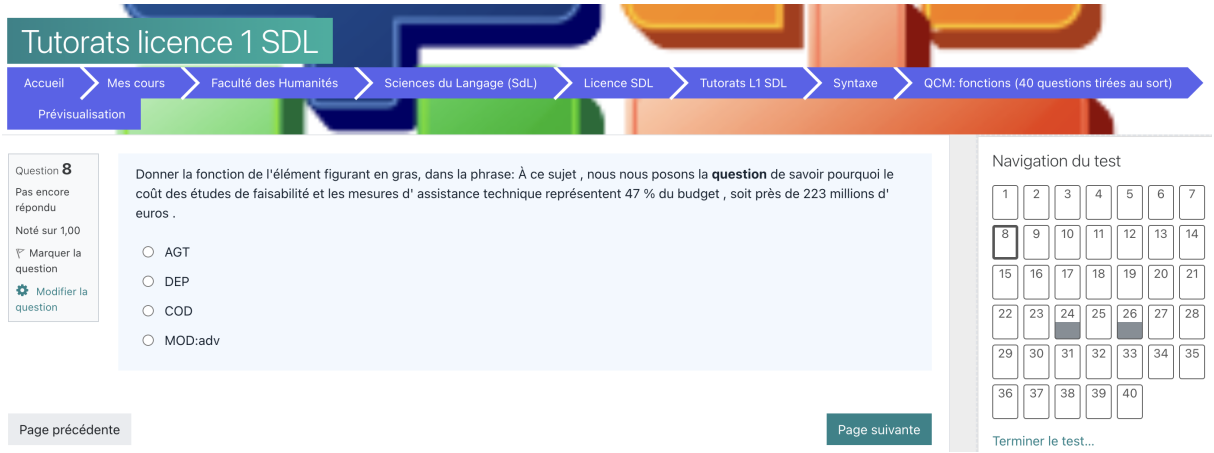

Figure 4: A quiz on syntactic functions generated from a UD corpus

In the examples above, all quiz activities are essentially text files, structured with the GIFT format. As such, the generated questions are ready to import into a Moodle question database to be used either as a formative or as an evaluative assessment activity.

3 First results: automatic syntax quizzes in the classroom

We first introduced automatically-generated syntax quizzes to groups of L1 students enrolled in the linguistics curriculum at our university in 2018. Initially, the aim was essentially to test different Moodle activities, while still retaining a classical "chalk-and-talk" approach. The COVID-19 pandemic, and the lockdowns that followed, have forced us to transform what was initially a mere addition to a classical teaching plan into our main formative and evaluative assessment tool. With a total of over 200 L1 students

of the budget, or nearly 223 million euros.

in 2019-2020, we had to devise a workable corpus-to-quiz solution that would provide large amounts of relevant formative and evaluative activities throughout a whole academic year. We kept using the material developed during that period even after lockdowns were lifted, and we are happy to report that, after having exposed over 800 L1 students to our automatically-generated quizzes over the course of four years, the basic concept can be validated. Students generally find it reassuring to be able to train themselves on large sets of syntax quizzes in preparation for mid-term and end-of-term exams. The fact that Moodle can provide an instant feedback on their performance is a clear motivation and engagement booster, as opposed to traditional syntax exercises. From the instructor's point of view, learner analytics processed by Moodle make it possible to easily identify "hard" or "easy" questions post-hoc, in order to fine-tune our growing set of syntax quizzes. We are now contemplating how to integrate the generated quizzes into other LMS platforms, and how to devise an interactive electronic syntax textbook, by using Jupyter books in conjunction with Moodle and other LMS platforms.

References

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. *Treebanks: Building and using parsed corpora*, 165–187.

Bick, E. (2001). The VISL System: Research and applicative aspects of IT-based learning. *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*.

Bick, E. (2004). Grammar for fun: IT-based grammar learning with VISL. *Copenhagen studies in language*, *30*, 49.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., & de La Clergerie, É. V. (2014). Deep Syntax Annotation of the Sequoia French Treebank. *International Conference on Language Resources and Evaluation (LREC)*.

Colin, É. (2020). *Traitement automatique des langues et génération automatique d'exercices de grammaire* (Doctoral dissertation). Université de Lorraine.

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, *47*(2), 255–308.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2008). The Sketch Engine. *Practical Lexicography: a reader*, 297–306.

Perez-Beltrachini, L., Gardent, C., & Kruszewski, G. (2012). Generating Grammar Exercises. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 147–156.

Uibo, H., & Bick, E. (2005). Treebank-based research and e-learning of Estonian syntax. *Proceedings of Second Baltic Conference on Human Language Technologies: Second Baltic Conference on Human Language Technologies*, 4–5.

Wijlff, A. (2006). VISL in Danish schools. *English Teaching: Practice & Critique (University of Waikato)*, *5*(1).