Hard Time Parsing Questions: Building a QuestionBank for French

Djamé Seddah[⋆ ⋄] **Marie Candito**[⋆ ◦]
[⋆]Alpage, INRIA [⋄] Université Paris-Sorbonne [◦]Université Paris Diderot
djame.seddah@paris-sorbonne.fr marie.candito@linguist.univ-paris-diderot.fr

Abstract

We present the French Question Bank, a treebank of 2600 questions. We show that classical parsing model performance drop while the inclusion of this data set is highly beneficial without harming the parsing of non-question data. when facing out-of-domain data with strong structural divergences. Two thirds being aligned with the English QuestionBank (Judge et al., 2006) and being freely available, this treebank will prove useful to build robust NLP systems.

Keywords: Treebanking; Parsing; Question-phrase

1. Introduction

One of the ever-recurring issues in statistical parsing is the matter of out-of-domain parsing. Namely how to make a parser trained on, by definition, a narrow domain able to cope with any kind of text. The range of possible issues can be characterized on a 3-axis graph where each axis denotes the divergence compared to an in-domain treebank, well edited, treebank: (i) a lexical divergence axis, (ii) a *noisy-ness* axis and (iii) a syntactic divergence axis. In this work, we focus on the syntactic divergences underlying the question genre. To do so, we present the French Question Bank (FQB), a French treebank of syntactically-annotated questions[1], and we investigate the performance of in-domain trained parsers on this data set, showing a clear loss of performance brought by structural divergences at the functional level. When added to the training set, large improvements are shown, demonstrating the usefulness of this new data set.

2. French questions typology

Our motivation was to improve statistical parsing performance on questions, which are crucial for e.g. QA and yet difficult to (statistically) parse due to the often non-canonical structure and word order they exhibit. More precisely, we can distinguish roughly the questions with an extracted (i.e. fronted) phrase from the *in situ* questions, which exhibit the canonical word order.

In situ questions can be split into: **(i)** those containing a *wh*-phrase, namely a constituent with embedded interrogative determiner, adjective, pronoun or adverb[2], but appearing in canonical position (for instance : *Paul a mangé quel dessert? lit. 'Paul has eaten which dessert?' (Which dessert did Paul eat?)*, or **(ii)** yes/no questions, for which the interrogative status is either marked by prosody / question mark only (*Paul a déjà mangé ? lit. 'Paul has already eaten?'*, or using a nominative clitic after the inflected verb. The clitic is either redundant (*clitic doubling*) with a non-anaphoric pre-verbal subject (*Paul a-t-il déjà mangé? 'Paul has-CL-NOM-3rd-sg already eaten?' (Has Paul already eaten?)*) or it is anaphoric and plays the role of the subject (*A-t-il déjà mangé? 'Has-CL-NOM-3rd-sg already eaten?' (Has he already eaten?)*).

Extracted wh- phrase Questions with an extracted wh-phrase show a more different word order/structure. We can distinguish:

case (1) Fronting, with pre-verbal subject and clitic doubling: *Quel dessert Paul a-t-il mangé? lit. 'Which dessert Paul has-CL-NOM-3rd-sg eaten?' (Which dessert has Paul eaten?)*

case (2) Fronting, with inverted non-clitic subject: *Quel dessert a mangé Paul* lit. *'Which dessert has eaten Paul?'*

case (3) Fronting with inverted clitic subject: *Quel dessert a-t-il mangé ? lit. 'Which dessert has-CL-NOM-3rd-sg eaten?' (Which dessert has he eaten?)*

While (2) can also appear in an embedded clause, the embedded equivalent of (1) is without clitic doubling.

Other syntactically-specific questions are the ones with a complex wh-marker *est-ce que*:

case (4) yes/no questions of the form *est-ce que + SENT*: *Est-ce que Paul a déjà mangé ? lit. 'Is-it that Paul has already eaten?' (Has Paul already eaten?)*

case (5) form *qui/qu' est-ce que/qui + SENT-with-gap*: *Qu'est-ce que Paul a mangé? lit. 'What is-it that Paul has eaten?' (What has Paul eaten?)*

case (6) form *qu' est-ce que + NP*: *Qu'est-ce que le platine?* lit. *'What is-it that platine' (What is platine)*

3. Questions in French corpora

We now focus on questions in the French typical corpora usable for training statistical parsers. The French treebank (FTB) (Abeillé and Barrier, 2004)) is the most used treebank for that purpose, being both the first and the biggest. Other treebanks were developped later, in particular some out-of-domain treebanks using the same annotation scheme : the SEQUOIA treebank (Candito and Seddah, 2012), a well-edited out-of-domain small treebank, and the FRENCH SOCIAL MEDIA BANK, FSMB (Seddah et al., 2012), which originates in web forums and social media content.

As already noted for English by Judge et al. (2006), questions are generally under-represented in treebanks. Indeed, this observation is confirmed the figures presented in Table

[1]This is to our knowledge the first non-English QuestionBank.

[2]The main ones are *qui (who), que (what), quel (which), quoi (what[-hum]), quand (when), où (where), comment (how), ...* Wh-words in French are sometimes called "mot-qu", as the "wh" French counterpart is "qu".

1: less than a few hundred sentences from the various cited treebanks do contain a *wh*-phrase.

	FTB-UC *(2007)*	FSMB *(2012)*	SEQUOIA *(2012)*	FQB *(-)*
# words	350947	20584	69356	23236
# sentences	12351	1656	3204	2289
Av. sent. length	28.41	12.42	21.64	10.15
# sentences with *wh*-phrase (%)	210 (1.68)	61 (3.68)	85 (2.65)	1710 (74.7)
# extracted *wh*-phrase				
wh- case 1	12	2	12	177
wh- case 2	22	3	27	800
wh- case 3	13	11	12	79
wh- case 4	0	2	0	1
wh- case 5	1	0	0	17
wh- case 6	0	0	4	134
# of *in situ wh*-	172	54	30	502

Table 1: French Treebanks statistics. Top: general statistics. Bottom: Number of *wh*- questions, broken down using the typology used in section 2.

Data Sources The raw questions have several origins: (i) the translation to French of the TREC 8-11 track test sets [3], (ii) the frequently asked questions section of various official French organization websites [4], (iii) and the question test set of the CLEF-03 Question-Answering shared task (Magnini et al., 2004) and (iv) questions from the Marmitton cooking web forums. All the first 3 blocks of questions are correctly edited, although the TREC part was lightly corrected to replace some strong Canadian-French idiosyncrasies by their standard French counterparts. We left the web forum questions unedited so that the difficulties of handling noisy questions can be correctly assessed.

SOURCE	# OF SENTENCES
TREC 08-11	1893
Faq GVT/NGOs	196
CLEF03	200
sub-total	*2289*
Web	285

Table 2: Source of FQB sentences.

The difficulties gathering question data in French entailed a relatively unbalanced corpus, compared for example to the Question Bank (QB) (Judge et al., 2006), as shown by the divergence in size between our corpus parts (see Table 2). Let us note that the TREC part of the French Question Bank (FQB) is aligned with the first 1893 sentences of the QB. Joining those resources could prove useful for the evaluation of some syntax based machine translation system if not for the bootstrapping of such systems.

[3] http://www-rali.iro.umontreal.ca/rali/?q=node/9

[4] Social Welfare (CAF), IRS (Trésors public), employment agency (Pôle Emploi), National Statistics Agency (INSEE), UNESCO

4. Annotation Scheme

In order to obtain evaluation treebanks compatible with parsers trained on the FTB, we have used as basis the FTB annotation scheme and followed as much as possible the corresponding annotation guidelines for morphology, phrase structure and functional annotation (Abeillé et al., 2003). More precisely, we started from a slight modification of this annotation scheme, referred to as the FTB-UC (Candito and Crabbé, 2009) and added specific guidelines for handling idiosyncrasies tied to question-phrase specificities.[5]

As far as grammatical function tags are concerned, we used an additional function label DIS for dislocated phrases. Such phrases appear either at the beginning or the end of a clause, and are coreferent with a (redundant) clitic appearing on the verb. It can occur in declarative sentences (e.g. *Paul les a mangées, les fraises* lit. *'Paul CL-ACC-pl has eaten, the strawberries' (Paul has eaten the strawberries)*, but in the FQB it appears massively in questions of the form *Qu'est-ce que NP* whose parse is shown in Figure 1 (cf. case 6 listed in section 2.).

In order to prepare a further deep syntax annotation layer, we also annotated all long distance dependencies using functional paths, following, among others, (Schluter and van Genabith, 2008; Chrupała, 2008). The motivation lies in the need to closely follow the FTB annotation scheme, therefore avoiding empty elements and traces. Other modifications such as assigning function labels to pre-terminal and participle phrases were applied so that a dependency conversion will be less sensitive to structural ambiguities than the original conversion developed by Candito et al. (2010a).

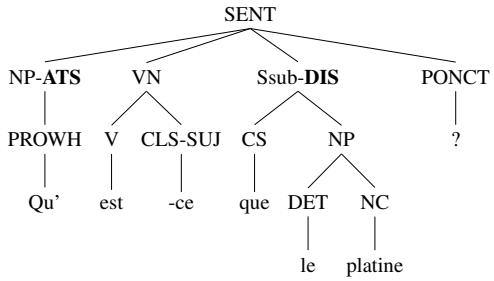

Figure 1: Dislocated example for lit. *What is-it that platine? (What is platine?)*

4.1. Annotation Methodology and Evaluation

We followed the same annotation protocol as (Candito and Seddah, 2012). Namely, two annotators working on the output of two parsers (the Berkeley parser (Petrov et al., 2006) and the first-phase parser of Charniak (2000)) fed with gold input (generated from a previous annotation phase). Resulting corrected parses were then adjudicated.

To assess the quality of annotation, we calculated the inter-annotator agreement using the Parseval F-measure metric between two functionally annotated set of parses (Table 3).

[5] Should this paper be accepted, we will provide more details on the annotation scheme.

We note that our agreement scores are higher than those reported in other out-of-domain initiatives for French (Candito and Seddah, 2012; Seddah et al., 2012). This can be due to the smaller average sentence length of the FQB, and to the fact that the annotators were already trained for the task. [6]

A vs B	A vs Gold	B vs Gold
97.54	95.72	97.21

Table 3: Inter-annotator agreement

5. *Parsability* of the FQB

As we said earlier, the motivation behind this work is to extend the French treebanks with more questions in order to bring more robustness to treebank-based parsers. In the absence of such data set, there is no visibility of the performance to expect from currently available parsers for French on questions. In this section, we present an overview of off-the-shelf parsers, using their widely available trained models. To evaluate constituency parsing, we used the Petrov et al. (2006) parser (BKY) with the baseline grammar extracted by Candito and Crabbé (2009), and the MALT parser (Nivre et al., 2006) with its already available French model (Candito et al., 2010b) coupled with the MELT tagger (Denis and Sagot, 2009). Both were trained on the canonical FTB training set. We therefore removed all FQB annotation scheme extensions (making the task obviously a bit easier). We also converted the BKY output to dependencies following Candito et al. (2010b).
As we did not perform any tuning and only provide baseline results, by lack of space, we report only results on the canonical FTB test set and on the non-web part of the FQB.
Table 4 presents surprisingly high results (F1 of 83.85% for the FTB, 81.67 for the FQB, with Bky's internal tagging). The reason comes from the sentence length distribution, with more than 99% of its sentences containing less than 20 words. On these shorter sentences, performance gap between in-domain and out-domain data is more perceptible (88.07 (FTB) vs 84.16 (FQB)), even though the FTB subset contains much less sentences (380 vs 1235 initially). As the FQB contains more than 13% of out-of-vocabulary words, the use of gold part-of-speech improves the overall performance by 4 points.

When analyzing the obtained parses, we could notice that such a high performance on out-of-domain data is a direct consequence of using a phrase-based metric that does not take grammatical functions into account. On a non configurational treebank such as the FTB where the difference between arguments and adjuncts is made at the functional level (no VP node), evaluating raw parses of questions, with

[6]The main difficulties we experienced lied in the difficulty to annotate complex named entities such as movie titles. The solution we choose (a proper structure) is not fully satisfying in the absence of quotes, or upper case letters (eg. "Who saw who framed Roger Rabbit? ").

	FQB		FTB	
POS	none	gold	none	gold
	Bracketing Fmeasure (all sent.)			
w/o funct	83.85	86.09	81.67	83.50
with funct	65.21	69.90	74.4	76.06
	Bracketing Fmeasure (≤20 sent)			
w/o funct	84.16	86.40	88.07	90.48
with funct	65.43	69.87	78.84	80.91
	Pos accuracy (all sent.)			
	92.05	98.98	97.29	99.93

Table 4: Baseline phrase-based results (BKY).

frequent subject-verb inversion, makes very little sense. This is confirmed by keeping function labels for the evaluation, which shows a reversed situation (the drop in performance shown in Table 4 is approx. 20 points for the FQB, and only 10 for the FTB).
Studying dependency-based parsers' results leads to less contrasted observations where Malt parser slightly outperforms phrase-based conversion in predicted tagging mode, while the opposite is verified for BKY in gold mode. One explication could come from the fact that the tagger associated with Malt makes use of a lexicon to handle OOVs, while BKY does not. In all cases, the performance of both parsers on this data set stands behind the state-of-the-art in FTB parsing by a significant margin (10 points), Candito et al. (2010b) report 86.2 for a Malt baseline on the FTB test set). Interestingly, unlabeled attachment scores on the FQB are on-par with previous results.

	FQB		FTB	
	LAS	UAS	LAS	UAS
(all sent)				
BKY *(own tagging)*	76.22	86.68	83.89	87.22
Malt *(Tagger)*	76.48	87.70	81.50	84.98
BKY *(Gold)*	81.48	92.11	85.91	88.95
Malt *(Gold)*	80.84	92.22	83.53	86.47
(≤20)				
BKY *(own tagging)*	76.05	86.77	86.80	90.36
Malt *(Tagger)*	76.40	87.88	86.26	89.73
BKY *(Gold)*	81.43	92.35	89.81	93.05
Malt *(Gold)*	80.70	92.43	88.86	91.88

Table 5: Baseline Dependency Results (Malt vs BKY – const. to dep.

Space is missing for an in-depth error analysis but we can hypothesis that structural differences between the FTB phrase-based annotation scheme and the FQB led to different labeling schemes but somewhat not in term of governing schemes. This suggests, as expected, that the inclusion of question data to the FTB would boost parsing performance.
Indeed, we carried out a 10 fold cross-validation experiment with our phrase-based architecture (BKY own tagging, constituent to dependency conversion) where 90% of

the FQB was added to the FTB training set in each fold. Results Table 6 show a drastic improvement compared to our baseline. Note that this gain does not only originate from the POS accuracy gain (97.51 vs 92.05) as all our parsing scores are higher in the realistic cross-validation mode than they were in gold POS mode with the sole FTB for training. A backtest of each model on the FTB test section delivers an averaged F1 score of 82.14% (no POS given), slighter higher than the 81.67% baseline.

	FTB +FQB			
POS	none	gold	none	gold
	(all sent.)		*(≤20 sent.)*	
LAS	85.51	87.34	85.71	87.49
UAS	94.41	96.03	94.84	96.4
FMeasure	93.33	94.6	93.95	95.11
Pos	97.51	99.4	-	-

Table 6: Cross-validation experiments using the FTB and the FQB

Conclusion

We introduced the first QuestionBank outside English, bringing a new genre to the existing French data set. Because statistical parsing models are notoriously biased toward the domain of their training model, the availability of a treebank made of questions for French will help building more robust parsers, useful for example in syntax-augmented question answering system. However, we showed in this work how this data set could be used to close the question genre out-of-domain gap. Once more unlabeled question data are made available for French, complementary techniques, such as uptraining (Petrov et al., 2010) or paraphrasing (Choe and McClosky, 2015), will help to further improve question parsing for French.

A large part of the FQB being aligned with the QB (Judge et al., 2006), this treebank will pave the way for cross-linguistics work. The FQB is freely available at `http://alpage.inria.fr/FrenchQuestionBank`.

Acknowledgment

We thanks our anonymous reviewers for their comments. We are grateful to Benoit Crabbé for his remarks on a earlier version of this work and to Corentin Ribeyre for his generous help. This work was funded by the Program "Investissements d'avenir" managed by the *Agence Nationale de la Recherche* ANR-10-LABX-0083 (Labex EFL).

References

Abeillé, A. and Barrier, N. (2004). Enriching a french treebank. In *Proc. of LREC'04*, Lisbon, Portugal.

Abeillé, A., Clément, L., and Toussenel, F., (2003). *Building a Treebank for French*. Kluwer, Dordrecht.

Candito, M. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of IWPT'09*, Paris, France, October.

Candito, M. and Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *In Proceedings of Traitement Automatique des Langues Naturelles (TALN 2012)*, Grenoble, France, (To appear).

Candito, M., Crabbé, B., and Denis, P. (2010a). Statistical french dependency parsing: Treebank conversion and first results. In *Proc. of LREC*.

Candito, M., Nivre, J., Denis, P., and Henestroza, E. (2010b). Benchmarking of statistical dependency parsers for french. In *Proc. of CoLing'10*, Beijing, China.

Charniak, E. (2000). A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139, Seattle, WA.

Choe, D. K. and McClosky, D. (2015). Parsing paraphrases with joint inference. In *In Proceedings of ALC-IJCNLP*. Association for Computational Linguistics.

Chrupała, G. (2008). *Towards a machine-learning architecture for lexical functional grammar parsing*. Ph.D. thesis, Dublin City University.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proc. of PACLIC*, Hong Kong, China.

Judge, J., Cahill, A., and van Genabith, J. (2006). QuestionBank: Creating a Corpus of Parse-Annotated Questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Anual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 497–504, Sydney, Australia.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., and de Rijke, M. (2004). Creating the disequa corpus: a test set for multilingual question answering. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 487–500. Springer.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July.

Petrov, S., Chang, P., Ringgaard, M., and Alshawi, H. (2010). Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713. Association for Computational Linguistics.

Schluter, N. and van Genabith, J. (2008). Treebank-based acquisition of lfg parsing resources for french. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The french social media bank: a

treebank of noisy user generated content. In *Proceedings of CoLing'12*, Mumbai, India.