From Noisy Questions to Minecraft Texts: Annotation Challenges in Extreme Syntax Scenarios

Héctor Alonso Martínez, Djamé Seddah, Benoît Sagot

From Noisy Questions to Minecraft Texts: Annotation Challenges in *Extreme Syntax* Scenarios

Héctor Martínez Alonso[1] **Djamé Seddah**[1,2] **Benoît Sagot**[1]

(1) Inria Alpage & Université Paris Diderot (2) Université Paris Sorbonne

{firstname.lastname}@inria.fr

Abstract

User-generated content presents many challenges for its automatic processing. While many of them do come from out-of-vocabulary effects, others spawn from different linguistic phenomena such as unusual syntax. In this work we present a French three-domain data set made up of question headlines from a cooking forum, game chat logs and associated forums from two popular online games (MINECRAFT & LEAGUE OF LEGENDS). We chose these domains because they encompass different degrees of lexical and syntactic compliance with canonical language. We conduct an automatic and manual evaluation of the difficulties of processing these domains for part-of-speech prediction, and introduce a pilot study to determine whether dependency analysis lends itself well to annotate these data. We also discuss the development cost of our data set.

1 Introduction

The continuous growth of the volume of user-generated content (UGC) published on the web stresses the need for efficient way to automatically process this type of data. Yet not only the volume of UGC increases; it also becomes increasingly varied, resulting in the need for domain- and register-adaptation methods and resources for processing UGC in all its diversity.

In this work, we present a feasibility study on dependency syntax annotation for three UGC domains in French, namely a cooking forum, in-game chat logs, and associated gaming forums. While these data sources are very different, they share the characteristic that their content was produced within time or space constraints. Such constraints force the users to resort to a variety of linguistic strategies to efficiently convey their message.

The work described here shows that, on top of the well-known problem of out-of-vocabulary words, automatic annotation and processing of UGC presents a double challenge. First, in order to interpret most of the data , it is crucial to take into account the interplay between **context and domain knowledge** on the one hand and their linguistic impact. This is because most messages can only be fully analysed with a good knowledge of the domain and context at hand. For instance, in-game chat logs can only be understood with a knowledge of the video game being played and of many game-specific terms, a representation of the game situation when a chat message is written and of, as well as a model of the ongoing dialogue, as such data is conversational by nature. Also, time- or space-constrained writing favors **fragmentary writing** that is more prone to ellipses, which makes linguistic analysis, especially parsing, more difficult. In addition to this highly contextual nature, the many idiosyncrasies plaguing UGC and make its analysis more challenging than regular out-of-domain text force most morpho-syntactic processing to be *extremely* robust at all levels of analysis.

In Section 3 we describe the data collection process, and give a first quantiative description of how our datasets are different from standard datasets. Two of our datasets were already annotated, and we manually annotated the third one. In Section 4 we provide a threefold categorisation of lexical variation in UGC. Finally, Section 5 is dedicated to our feasibility study regarding dependency annotation of our data using the Universal Dependencies annotation scheme. It also includes a brief discussion about annotation costs, an issue rarely explicitely discussed.

Our contribution is threefold: (i) an empirical account of the phenomena behind domain-shift performance drops in French UGC data processing, (ii) a syntactic study on the applicability of Universal Dependencies to French UGC, and (iii) the first corpus obtained from MINECRAFT and LEAGUE OF LEGENDS gaming logs. All corpora and annotations are freely available.

Proceedings of the 2nd Workshop on Noisy User-generated Text,
pages 127–137, Osaka, Japan, December 11 2016.

2 Related Work

Before the global availability of social-media feeds, studies on the difficulties of out-of-domain statistical parsing have been focusing mainly on slightly different newspaper texts (Gildea, 2001; McClosky et al., 2006b; McClosky et al., 2006a), biomedical data (Lease and Charniak, 2005; McClosky and Charniak, 2008) or balanced corpora mixing different genres (Foster et al., 2007).

For such data, which is as edited as standard data sources, the problem is "simply" a matter of domain adaptation. It is far from being the case for UGC data, as shown by Foster (2010). Indeed, in her seminal work on parsing web data, different issues preventing reasonably good parsing performance were highlighted; most of them were tied to lexical differences (coming from either genuine unknown words, typographical divergences, bad segmentation, etc.) or syntactic structures absent from training data (imperative usage, direct discourse, slang, etc.). This suboptimal parsing behavior on web data was in turn confirmed in follow-up works on Twitter and IRC chat (Foster et al., 2011a; Gimpel et al., 2010; Elsner and Charniak, 2011). They were again confirmed during the SANCL shared task, organised by Google, aimed at assessing the performances of parsers on various genres of Web texts (Petrov and McDonald, 2012). Foster (2010) and Foster et al. (2011b) noted that simple lexical and tokenisation convention adaptation to the Wall-Street Journal text genre could increase the parsing performance by a large margin. In addition, Seddah et al. (2012) showed that a certain amount of normalisation brought a large improvement in POS tagger performance of French social media texts. These normalisation steps mostly apply at the lexical level, at the very definition of what constitutes a minimal unit. Plank et al. (2014) attempt to quantify how much of the domain-specific variation of POS labeling is a result of different interpretations, and how much is arguably just noise.

Regarding the study of French UGC, our starting point is the part-of-speech and phrase-structure annotation guidelines by Seddah et al. (2012). However, we conduct our syntactic analysis in terms of dependency structures.

3 Data Collection and Part-of-Speech Annotation

Our dataset contains three different sources of user-generated content. Two of them are logs of multi-player video-game chat sessions, MINECRAFT and LEAGUE OF LEGENDS[1], the last one is made of instant cooking-related web questions from MARMITON[2], a widely popular French recipe website. This set of questions was collected during the building of the French QuestionBank (Seddah and Candito, 2016) but was not described nor analysed because of its syntactic peculiarity and was thus considered by the authors as a clear outlier. We chose to include this sample in our study because it offers a sharp contrast with video games chat logs in term of domain variation while retaining a *live* nature: users asks questions related to their immediate needs and expect a quick answer.

The LEAGUE OF LEGENDS data set was collected by Lamy (2015) in early 2015 and consists of two types of recorded user interactions: a first part is the record of discussions occurring during an on-going game session while the second part consists of different players post-game discussion that took place both on official game forums and on unofficial boards.

Table 1 presents the corpus properties. The figures highlight its unbalanced nature and confirm the weight of the medium of appearance, be it a "crude" chat system, tweets or a classic web forum. Both live in-game chat sessions (MINECRAFT and LEAGUE OF LEGENDS *in-game*) display similar properties in term of small average length, while the MARMITON and LEAGUE OF LEGENDS *outside* data exhibit rather different properties: sentences are either smaller with a small standard deviation for the former or on average longer with a strong variation for the latter.

3.1 Measuring the "Non-Canonical-ness" of our Data

In order to quantitatively corroborate our intuitions concerning the level of noise in our corpora, and for measuring their various levels of divergence compared to the French Treebank (Abeillé et al., 2003), we used an *ad hoc* noisiness metric. It is simply defined as a variant of the Kullback–Leibler (KL)

[1]Resp. `http://www.minecraft.com` and `http://leagueoflegends.com`

[2]`http://www.marmitton.org`

	# OF SENTENCES	# OF TOKENS	AV. LENGHT	STD DEVIATION	*noisiness* LEVEL (KL)
MARMITON	285	2080	7.30	2.57	3.43
LEAGUE OF LEGENDS	453	5106	11.27	12.55	3.48
in-game	254	961	3.78	2.95	2.98
outside	199	4145	20.82	13.57	3.46
MINECRAFT	236	913	3.87	3.94	3.10
ALL	974	8099	8.32	9.38	3.58

Table 1: Corpus properties.

divergence[3] between the distribution of trigrams of characters in a given corpus compared to a reference. Unlike the raw text properties, the KL divergences we computed (Table 1) range around the same figures (between 2.98 and 3.50). These levels of *noisiness* appear much higher than the ones reported by Seddah et al. (2012) for more classical source of user-generated content. This discrepancy can be caused by two factors: either the texts themselves contain a lot of noise and depart strongly from the writing norm, or their lexical domain is so different from the French Treebank's one that the variation it carries can be considered as noise. If we compare the KL divergence between our 3 data sources (Table 2), we can see that the trigrams distributions are somehow "closer" to each other than they are to edited text.

A / B	A vs B	B vs A
MARMITON / LEAGUE OF LEGENDS	1.36	0.81
MARMITON / MINECRAFT	1.88	1.38
LEAGUE OF LEGENDS / MINECRAFT	1.30	1.40

Table 2: *noisiness* levels for each of our sub-corpora.

3.2 PoS Annotation of Additional Data

As we mentioned in the introduction, our work takes its source in a will to build robust parsing models regardless of the domain variation. The challenge here is that variation may cover anything from historical to biomedical texts, Twitter's feeds subtitles, chat logs and so on. For some machine learning purists, anything not present in the training data is purely an out-of-domain instance and coping with this variation should be left only to the model. In this point of view, the loss of performance can be circumvented by adding more data, annotated or not, thus enlarging the training set. However, in the case of UGC, the very notion of adding more data can be problematic: UGC covers basically everything that can be produced by someone with an internet access and at least some notion of a writing system. The need for more insights in terms of what to expect is becoming a striking issue.

This is why in order to allow further exploration of the language divergence at stake in our UGC sample data set and because the MARMITON and LEAGUE OF LEGENDS sub corpora were already annotated using the extended FTB tagset used for the French Social Media Bank, we annotated the MINECRAFT data set with the same type of annotations. This was done by one expert annotator. We also checked the consistency of the LEAGUE OF LEGENDS subset and corrected a few obvious errors. Let us add that this work is part of a pilot study investigating the development of a much larger data set, made of video games interaction, the annotations are then likely to evolve.

3.3 Exploring our Data: Automatic PoS Annotation Results

To provide a first glance on the difference between our three domains and the training data from FTB, we provide in Table 3 the PoS tagging accuracy results obtained with the PoS tagger *MElt* (Denis and Sagot, 2012) trained on edited journalistic data. Results are on-par with previously reported results on French

[3]It differs from a standard Kullback–Leibler distance because we apply a preliminary pre-processing to the corpora involved: (i) URLs, e-mail addresses and such, are removed, (ii) all characters that do not belong to an extensive list of characters that might be used in French sentences are replaced by a unique "non-standard character,"

		Baseline (FTB trained)		FTB trained+ Normalisation	
	OOV(%)	All	Unseen	All	Unseen
Marmiton	27.29	81.84	70.82	83.15	75.44
League of Legends	29.21	80.02	**52.92**	80.35	**45.77**
in-game chat	61.81	58.79	47.46	55.25	40.40
off-game session	21.64	84.95	56.41	86.13	60.42
Minecraft	52.57	**53.12**	**28.13**	58.27	36.04
all	31.36	77.44	52.19	78.62	45.42
French Social Media Bank (dev)	23.40	80.64	-	84.72	-
FTB (dev)	5.20	97.42	-	97.42	-

Table 3: POS tagging results using *MElt* trained on the French Treebank with and without normalisation. The tagger (Denis and Sagot, 2012) was trained on the canonical training section of the French Treebank (Abeillé et al., 2003) instance, FTB-UC, from (Candito and Crabbé, 2009). We used an extended version of the rewriting rules used to pre-annotate the French Social Media Bank (Seddah et al., 2012). They work jointly with the tagger to provide internal cleaned versions of a token, or a sequence of, which are tagged separately. Resulting POS tags are finally merged to the original token(s). (*e.g.wanna → want/VB to/TO → wanna/VB+TO).*

UGC tagging (Seddah et al., 2012; Nooralahzadeh et al., 2014): normalisation helps to cover some of the most frequent lexical variations and hence improves substantially the tagging accuracy. However in the case of *in-game* League of Legends chat session, the normalisation is detrimental to the overall tagging performance as well as for unseen words. One obvious hypothesis is simply that the rules are applied deterministically and assign wrong PoSs while letting the pure tagging model work alone provide reasonable assumption on what would be the correct label for an out-of-domain word. Let us add that the *MElt* tagger makes a heavy use of features extracted from wide coverage lexicon, this lexicon itself adds a domain bias in case of known words used in a totally different syntactic context (which, according to the FTB guidelines, implies a different category). While Marmiton and League of Legends *off-games* tagging results are in the same range than the FTB ones, Marmiton results exhibit vastly inferior performance but seem to benefit from a normalisation step.[4] To summarise, in-game live session logs are harder to process than "regular" user-generated content.

However, a difference in a single metric such as POS accuracy or Kullback-Lubler divergence offers little information on the causes of such difference. In other words, each domain has its idiosyncrasies, and a domain shift is the result of the Cartesian product of the idiosyncrasies of source and target data.

The particularities of UGC are unbound, each new data source—defined by user demographics, user scenario, technical constraints and communication purpose—can spawn a myriad of idiosyncrasies that are beyond most domain-adaptation techniques that depend on selective sampling, self-training or other semi-supervised learning techniques. Therefore, the purpose of the present work is not only to potentially provide a new dataset to be used as additional training data for domain adaptation. Rather, we provide a close inspection of the main causes behind the expectable performance drops in tagging and parsing.

We therefore conduct a series of automatic and manual inspections to better understand the linguistic phenomena behind UGC linguistic variability. We explore the relation between predicting performance and annotation difficulty, which is seldom explicitly addressed (Plank et al., 2015).

4 A Threefold Categorisation for UGC Idiosyncrasies

Even though user-generated content does not constitute a uniform genre, many works have characterised its idiosyncrasies (Foster, 2010; Gimpel et al., 2011; Seddah et al., 2012; Eisenstein, 2013), which can be characterised on three axes, defined by the intentionality or communication needs of the word variants:

1. **Encoding simplification:** This axis covers ergographic phenomena, *i.e.,*phenomena aiming at reducing the writing effort, perceived as first glance as genuine misspell errors, and transverse phe-

[4]In fact, in the case of League of Legends in-game data, the normalisation step adds a significant amount of noise. A solution to this problem and more generally to the limitations of deterministic rule-based normalisation lies in the development of non-supervised or semi-supervised adaptative approaches.

nomena that include token contraction, "iwuz" for "I was" and over-splitting, "c t" instead of "c'était" (*it was*). Word types resulting form these categories pose challenges for their appropiate tokenisation, because they can either be split away by conventional tokenisers or retain conflated tokens, cf. Section 5.1.

2. **Sentiment expression:** This axis corresponds to marks of expressiveness, e.g., graphical stretching, replication of punctuation marks such as ???, emoticons, sometimes used as a verb such as Je t'<3 standing for Je t'aime (*I love you*). These phenomena aim at emulating sentiments expressed through prosody and gesture in direct interaction. Many of these symbols contain punctuation, that can lead to spurious tokenisation. Game logs have also a lot of platform-dependent particularities in the way they encode emoticons, e.g. the :smile: symbol, which is a placeholder to show a smile icon, is split as three different tokens (: smile :) by the tokeniser.
3. **Context dependency** This axis corresponds to the amount of context needed to understand a post. The nature of different user platforms will influence the domain knowledge necessary to understand the specific terms, from ingredients in MARMITON to weapon characteristics in LEAGUE OF LEGENDS. As in dialogue-based interaction, speech turns are marked by the thread structure and provide a context rich enough to allow varying level of ellipsis and anaphora. This structure can be superseded by hashtags (which example to add a meta data susceptible to attract attention ("likes"), providing a parallel source of information. Maybe more importantly, additional multimedia content can provides a rich source of context that we no dot have full access to when annotating, such as game state, sounds and images being displayed, etc.

5 Annotating Syntactic Dependencies

Before engaging in an annotation task, we conduct a feasibility study for the three domains at hand. While part-of-speech and syntactic annotation are not mutually independent, in this section we focus on how different typographic, lexical and syntactic phenomena can increase the difficulty of syntactic analysis. The previous work of (Kaljahi et al., 2015) has focused on phrase-structure parsing of forum text, which was made more difficult by the presence of grammatical errors. They report the benefits of correcting grammatical errors before conducting constituency parsing. Nevertheless, our corpora of choice are even more removed from the expectable newswire training data than theirs, and error correction as such is largely impossible.

Analysing noisy data using phrase structure might require postulating empty elements to be able to name non-terminals, or even provide a complete tree altogether. In contrast, dependency analysis reduces this problem by ensuring every token has a head, and that all edges form a tree. This alleged relative ease of annotation of dependency syntax over phrase structure led us to conduct our feasibility study using dependency.

As a dependency formalism, we use UD or Universal Dependencies (Nivre et al., 2016). In this formalism, there are no empty elements, and there is a preference for assigning headnedness to content words. For instance, for 'under the table', 'table' is the head of both the article and the preposition. We choose this formalism because we consider it more reliable to annotate our utterances when they are missing function words.

5.1 Tokenisation

UD defines dependency relations between *syntactic* words, and not between orthograpic words. To give account for this principe, orthographic words can be internally broken down in their forming syntactic words. In the general case of French, preposition-article contractions such as 'du' (*of the*, singular masculine) must be treated as two token for syntactic purposes, namely 'de' and 'le'. Beyond these few—but highly frequent—cases, French words do not need internal syntactic token analysis in the general case.

However, user-generated data has plenty of ergographic phenomena that blur the practical equivalence between tokens and syntactic words. For instance, the standard French expression"j'aime" (*I like*) gets split by the apostrophe into two tokens by any regular tokeniser, but dropping apostrophes that would

otherwise be obligatory is a common practice in user-generated data. Thus, 'jaime' would not be automaticaly analysable without using strategies for apostrophe recovery. While the UD formalism does not oblige us to correct every single typographical or grammatical error, we are obliged to do so when an orthographic word corresponds to more than one syntactic words. Table 4 provides some detailed examples of these conflated tokens and the syntactic tokens they contain in their UD represention.

Form	Standardised	POS	Translation
ta	tu as	PRON VERB or PRON AUX	*you have*
ya	il y a	PRON PRON VERB	*there is*
jaime	j'aime	PRON VERB	*I like*
o	au (à le)	ADP DET	*in the*

Table 4: Usual non-standard conflated tokens in UGC, along with their standardised form and UD part-of-speech for their syntactic tokens once analysed.

5.2 Effects on Syntax of Time and Space Limits

Many user-generated data sources share the common characteristic that they are written under time and space limits. Twitter and SMS messages present well-known case of spatial contraints, namely 140 and 160 characters respectively. The MARMITON data presents a more lenient spatial restriction, because the sentences we use are discussion forum header questions. In our game log data, we find that text has been written under time constraints. In particular on Minecraft, we observe shorter messages and a higher prevalence of contractions, ergographism and typos.

Besides the obvious effect on spelling, time and space limitations also influence syntactic choices, because they also foster dropping less-relevant elements like articles or punctuations, but more importantly also whole lexical items or spans. Indeed, user-generated data presents plenty of candidates can we can interpret as ellipsis. While the pragmatics of an internet cooking forum can be very idiosyncratic, even newswire has a frequent source of syntactic anomalies, namely headlines (Perfetti et al., 1987). Headlinese is also a result of writing under space constraints, and is a particular use of language that is very prone to ellipses, e.g the 'U.S. Futures Higher Ahead of GDP Data'[5] has no main verb and only one function word. We expect the headline-esque, conversation-starting MARMITON questions to present ellipsis and disfluences resulting from its spatial limitations and its almost oral nature.

Finally, in the scenario of syntax under space and time contraints, we also observe that many relations between clauses are not explicitly marked. We also give account for the cases of parataxis in our data.

5.3 Code-Switching and Direct Speech

The two game-log domains contain plenty of English game terms, and we observe quite a lot of code-switching. When only a token is code-switched, it's potential difficulties of analysis can be resolved during part-of-speech annotation. In this aspect, an atomic code-switch does not differ from using a loanword (Example 1). When the code-switching spans more than one token, difficulties arise, as a certain span can have its own internal syntactic structure in English (Example 2), while being embedded in a French clause. Similarly, we also observe examples of direct speech with an embedded full clause in English (Example 3).

1. LoL: J' adore jouer Elise mais faut la **up** merci .
   I love playing Elise but must up her thanks .
2. Minecraft: qui a **stuff** ? **for me**
   who has stuff ? for me
3. LoL: J' ai donc dit d' un ton désperé en **all "GG guys, you were better , can' t carry this:** "
   I have thus said in a desperate tone in all "GG guys, you were better, can' t carry this: "

For Examples 2 and 3, it is necessary devise strategies that allow a full analysis of the main language, and potentially provide an analysis for the code-switched language in an alternative layer.

[5]The Wall Street Journal, online edition, October 28, 2016.

5.4 Quantitative Analysis

We have taken a sample of 100 unique sentences with at least 4 tokens from each of the three domains. One of the authors of this article, experienced in dependency annotation, evaluated the difficulties in annotating them using UD dependencies. Sentences are shorter in Minecraft, and only retrieved 87. Notice that we do not conduct any sentence splitting, and we treat each of the statements entered by the users as complete utterances. While some of these utterances could be split in clauses, other sequences of utterances could be joined to form a single sentence. However, we believe that treating each statement as a complete utterance is the most realistic scenario for later automatic processing.

Table 5 presents the percentage of sentences in each sample that present annotation difficulties in our UD pilot. From the phenomena described in this section, we have determined a series of categories, such as missing main verb (NoVerb), conflicting candidates to main predicate (Pred), parataxis (Parat), code-switching (CoSwi), missing punctuation that harms understanding (Punct), tokenisation problems such as fragments or overzealous tokenisation (Tok), and cases of conflated tokens such as the ones presented in Section 5.1.

Domain	NoVerb	Pred	Parat	CoSwi	Punct	Tok	Conf
LEAGUE OF LEGENDS	3	3	17	39	10	8	0
MARMITON	42	7	2	0	0	2	11
MINECRAFT	16	1	14	17	15	8	31

Table 5: Corpus-wise percentages of annotation difficulties.

As expected, the two gaming corpora are more similar between themselves than to MARMITON, which presents a defining particularity in that 42% of the examples do not have a main verb. In this domain, users rely heavily on domain knownledge, using a language that is often very close to spoken French, such as “Steak , pâtes et ???” (*“Steak, pasta and ???”*), where the last conjunct is ommited in order to formulate a question. While we only find two examples of parataxis in MARMITON, we observe 17 and 14 in LoL and Minecraft. These two domains respond to more real-time data, and the relation between clauses is less often made explicit. In addition, most statements in MARMITON are monoclausal.

Code-switching and heavy loanword and in-game vocabulary are defining traits for the gaming corpora, which LoL being the corpus with more code-switching and more instances of direct speech. The sentences with these phenomena present the difficulty of how to represent both the contribution to the main French sentence, and the internal structure of the cited or code-switched span.

5.5 Annotated Examples

We have conducted some preliminary full dependency annotations for examples we considered illustrative of the different domains and their characteristic syntactic phenomena. Figures 1 and 2 shows three examples of non-trivial analyses, one for each domain. We provide glosses for the examples in the figure, and their translations are: A: “Which wine with a pig roast?", B: "But we have to think: assault tank" and C: “Every time 3VS1 and suddenly -2 P4".

Example A is a MARMITON example, which has no main verb. However, the main term of the question is “wine”(*wine*), and we treat is as the main predicate. Example B presents an ambiguity in main predicate choice. While “penser”(*think*) is the only verb of the sentence and makes a good candidate for main predicate, we can also see the whole span before the colon it as an extrapredicative, with “char”(*tank*) being the head of the predication.

Example C presents a series of idiosyncrasies. First of all, we had to treat the contractions “du" (*of the*, singular) and “des" (*of the*, plural) as two separate tokens, marked in parentheses following UD tokenisation principles, as well as an unusual typo,i.e. "couˆ" should be “coup”, literally *hit*, a formant of the expression “du coup”, which means *literally*. This last multiword, treated with a flat structure with the *mwe* label, is very common in spoken French, but not present in the French UD treebank.

More importantly, C shows an attachment ambiguity caused by part-of-speech ambiguity and verb ellipsis. A natural ellipsis recovery of example C would read as "Every time **there are** 3VS1, and

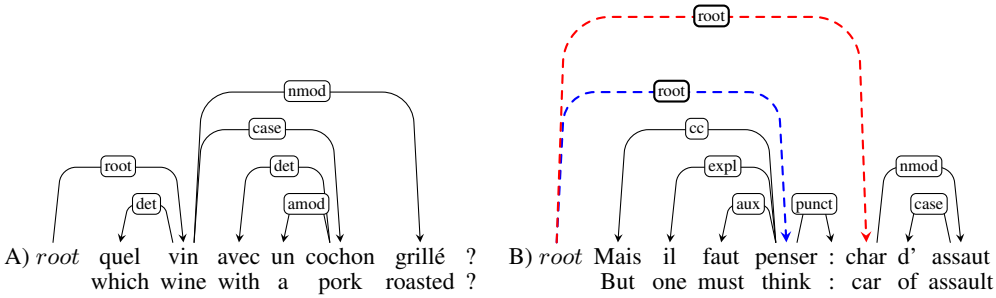

Figure 1: Example A, lacking a main verb, and B, with two contesting candidates for main predicated marked with dashed edges.

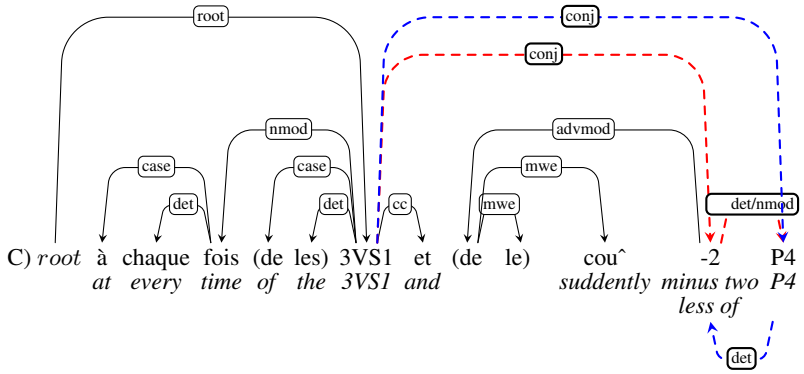

Figure 2: Example C, with two contesting structures from two different readings of the token "'-2".

suddenly **I have** -2 P4". The word "3VS1" stands for "3 versus 1", namely an uneven combat setting, and 'P4" refers to the character's protection armor. The token "-2" admits more than one analysis. The first analysis is the simple reading as number, complementing the noun "P4". A second analysis treats "-2" as a trascription of "moins de" (less than, less of), which would be the preferred analysis in case the verb recovery held. This example shows the interplay between frequent ellipses, ergographic phenomena (Seddah et al., 2012) and the need for domain knowledge in user-generated data.

5.6 Feasibility of Dependency Annotation

As said above, elliptical structures are frequent in our corpora and are a direct consequence of the *live* nature of the medias we choose to study. This is why at the beginning of this work, we decided to annotate these ellipsis, breaking thus the French Treebank guidelines, because we pretentiously thought we could always remain neutral in our interpretation of the missing syntactic context while agreeing on it and that it would be actually useful. Figure 3 shows an example of such annotation, here prototyped in a phrase-based manner, following the FTB guidelines. In this example, the two VNs (Verbal Nucleus in the FTB terminology) were supposed to respectively stand for "J'ai"/*I have* and "dois-je"/*should I* but they could have stand for anything else (preposition, adverbial phrase) while the strictly transitive verb "congeler"/*to freeze* is expecting a direct-object (showed by a trace in the clitics position pointing to the preceding NP). In short, trying to annotate such missing constructions was akin to numerous, if not pointless, inferences and led us to consider a UD analysis where we only had to make *bananas* the main

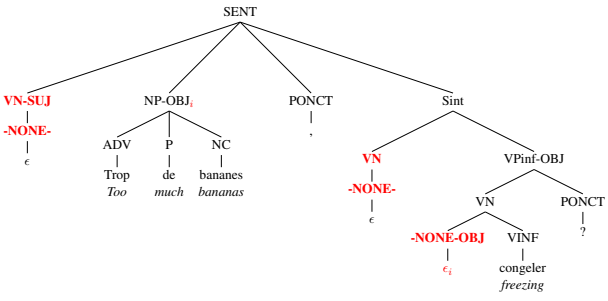

Figure 3: Prototype phrase-based analysis of an extreme elliptic case: *Too much bananas, freezing?*

predicate, and treat *freeze* as a subordinate clause.

Regardless of the difficulty of the domain, it appears that a UD dependency analysis lends itself to dependency annotation in an easier way: Since non-leaf relations appear between lexical words, this representation is more robust to missing determiners, prepositions and punctuations, even phrases. Also, if we used other dependency formalisms that for instance place prepositions as heads of nouns, it would be more difficult to annotate as it is the case using the current French Treebank dependency scheme (Candito et al., 2010). Nevertheless, dependency analyses conflate functional and structural information (Silveira and Manning, 2015), and some of the structural information can be lost in cases such as the Example C, discussed above.

Annotating dependencies lends itself well to noisy user-generated data. In a strict lexicalist analysis such as UD, where there are no tokens for unobserved words (e.g. dropped subjects, missing main verbs), we must build a structure from the existing words, and not from idealised sentence representations. We finally observe that for UGC, shorter sentences are harder to annotate. Indeed, sentences closer to the lower threshold of 4 tokens we have determined, seem to present more ellipsis, while longer sentences in our data have structures closer to more canonical syntax.

5.7 The Unspoken Costs of Treebank Annotation

As we all know, creating annotated data is a rewarding task, extremely useful for evaluation as well as for building feature-rich supervised models. Yet, it is time consuming and as generally said, relatively costly (Schneider, 2015) even though crowd-sourcing solutions through games with a purpose start to emerge (Guillaume et al., 2016). The dataset we presented in that paper are part of process that was initiated 5 years ago when we were confronted to the lack of syntactically-annotated out-of-domain dataset for French. The purely syntactic annotation phase for the LoL and Minecraft data is still ongoing and we expect it to be finished in the first few months of 2017. It is important to consider that such a task was made possible because of the experience we gained along the years and because we relied on a highly trained team of annotators. This training was the most important point in term of costs and had to be extended each time we added a new major annotation layer (from surface syntax to deep syntax for example, see (Candito et al., 2014) for details). Table 6 presents the costs of the treebanking effort (mainly led by Marie Candito and the second author) that was carried out by the INRIA's Alpage team since 2011 and led to the release of many out-of-domain and user-generated content treebanks. The figures do not include the costs of the permanent-position researchers involved in the design of the annotation scheme, in preliminary annotation, in the development of pre-annotation pipelines, post-annotation error detection and correction tools and in the training and supervision of the annotators. Considering that we annotated 4 different layers for about 7k out-of-domain sentences and 2 layers for 3.7k of UGC data, the average cost per sentence is about 3 euros, on par with what is known about the English Web Treebank (Bies et al., 2012) development costs (Mc Donald, PC). Unlike less costly initiatives that focused on gold standard or training data creation, our goal was also to provide a linguistic snapshot of a specific domain at a given time, useful as such for linguists.

	Start	Size	Morph.	Const.	Dep.	Deep Synt.[4]	Cost
		sent.		*man/month*			*euros*
Sequoia[1]	2011	3200	2	9	1	6	59k
FSMB 1[2]	2012	1700	1	2	-	-	13k
FSMB 2[2]	2014	2000	2	4	-	-	20k
FQB[3]	2014	2600	2	4	1	4	36k
LoL	2015	450	3	-	-	-	3k
Minecraft	2016	230	0.5	-	-	-	2k
		10180			41.5		133k

Table 6: Treebanking Cost at the Alpage team. *Morph.: morpho-syntactic annotation, Const: Phrase-based annotation, Dep: dependency conversion, Deep Synt: Deep syntax annotation.*

[1]:(Candito and Seddah, 2012),[2]:(Seddah et al., 2012),[3]:(Seddah and Candito, 2016),[4]:(Candito et al., 2014)

6 Conclusion

We have presented a three-corpus dataset of user-generated French data. We have shown how the deviation from conventional newswire data manifests in different ways, and not only lexical elements. For lexical elements, we have presented a threefold categorisation of UGC idiosyncracies.

We have subsequently conducted a pilot study to evaluate the a priori difficulties of applying a UD analysis on our data. We have determined that for our data set, after POS analysis, the difficulty of annotating depends on the amount of elided material and on the ability of the representation scheme to cope with the discrepancies between raw and syntactic tokens. While dependency syntax is more robust towards missing elements that phrase-structure representations, many the examples of the forum text presented cases of i.a. main-verb ellipsis, which are still difficult to tackle using dependency analysis. While the main problem for analysis of term-rich texts like those of gaming logs lies (LEAGUE OF LEGENDS and MINECRAFT) at the POS level, the *hidden* problem arises with perfectly standard text lacking most of the usual syntactic glue, including verbal predicates, which we find in MARMITON. For games, however, the other main problem lies in the strong contextualization of such production, making them barely understandable by a domain-outsider. Lastly, we believe our data can be used as a very interesting crash-test or a valuable reference material, so we make all data freely available at `http://alpage.inria.fr/Treebanks/ExtremeUGC/` :smile:

Acknowledgments

We thank our anonymous reviewers for their comments. This work was partly funded by the ANR SOSWEET and PARSITI projects, as well as by the Program "Investissements d'avenir" managed by the Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

References

Anne Abeillé, Lionel Clément, and François Toussenel, 2003. *Building a Treebank for French*. Kluwer, Dordrecht.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. Technical report, Linguistic Data Consortium,, Philadelphia, PA, USA.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *IWPT*, Paris, France.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *Proceedings of COLING 2010*, Beijing, China.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric Villemonte de La Clergerie. 2014. Deep syntax annotation of the sequoia french treebank. In *International Conference on Language Resources and Evaluation (LREC)*.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *HLT-NAACL*, Atlanta, USA.

M. Elsner and E. Charniak. 2011. Disentangling chat with local coherence models. In *ACL*.

J. Foster, J. Wagner, D. Seddah, and J. Van Genabith. 2007. Adapting wsj-trained parsers to the british national corpus using in-domain self-training. In *IWPT*.

J. Foster, O. Cetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan, and J. van Genabith. 2011a. #hardtoparse: Pos tagging and parsing the twitterverse. In *WOAM*.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011b. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *IJCNLP*, Chiang Mai, Thailand.

Jennifer Foster. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *NAACL*, Los Angeles, California.

Daniel Gildea. 2001. Corpus variation and parser performance. In *EMNLP*, Pittsburgh, USA.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith. 2010. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, School of Computer Science, Carnegie Mellon University.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL*, Portland, Oregon, USA.

Bruno Guillaume, Karen Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka (Japon).

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, and Joseph Le Roux. 2015. Foreebank: Syntactic analysis of customer support forums. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Laurine Lamy. 2015. Constitution d'un corpus d'évaluation web 2.0 fortement bruité et analyse syntaxique automatique préliminaire. Master's thesis, Université Paris Sorbonne (Paris IV), Paris, France, Septembre.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. *IJCNLP*.

David McClosky and Eugene Charniak. 2008. Self-training for biomedical parsing. In *ACL*, Columbus, Ohio.

D. McClosky, E. Charniak, and M. Johnson. 2006a. Reranking and self-training for parser adaptation. In *CoLing*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Effective self-training for parsing. In *NAACL*, New York City, USA.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Farhad Nooralahzadeh, Caroline Brun, and Claude Roux. 2014. Part of speech tagging for french social media data. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1764–1772, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Charles A Perfetti, Sylvia Beverly, Laura Bell, Kimberly Rodgers, and Robert Faux. 1987. Comprehending newspaper headlines. *Journal of Memory and Language*, 26(6):692–713.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *SANCL*, Montreal, Canada.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *ACL (2)*, pages 507–511.

Barbara Plank, Héctor Martínez Alonso, and Anders Søgaard. 2015. Non-canonical language is not harder to annotate than canonical language. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 148.

Nathan Schneider. 2015. What I've learned about annotating informal text (and why you shouldn't take my word for it). In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 152.

Djamé Seddah and Marie Candito. 2016. Hard time parsing questions: Building a questionbank for french. In *in Proceedings of LREC*.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *CoLing*, Mumbai, India.

Natalia Silveira and Christopher Manning. 2015. Does universal dependencies need a parsing representation? an investigation of english. *Depling 2015*, page 310.