The French Social Media Bank: a Treebank of Noisy User Generated Content

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, Vanessa Combet

The French Social Media Bank: a Treebank of Noisy User Generated Content

Djamé Seddah[1,2] Benoit Sagot[1] Marie Candito[1]
Virginie Mouilleron[1] Vanessa Combet[1]
(1) Alpage, Inria Paris-Rocquencourt & Université Paris 7
175 rue du Chevaleret, 75013 Paris, France
(2) Université Paris Sorbonne
7 rue Victor Cousin, 75006, Paris, France
firstname.lastname@inria.fr

ABSTRACT
In recent years, statistical parsers have reached high performance levels on well-edited texts. Domain adaptation techniques have improved parsing results on text genres differing from the journalistic data most parsers are trained on. However, such corpora usually comply with standard linguistic, spelling and typographic conventions. In the meantime, the emergence of Web 2.0 communication media has caused the apparition of new types of online textual data. Although valuable, e.g., in terms of data mining and sentiment analysis, such user-generated content rarely complies with standard conventions: they are *noisy*. This prevents most NLP tools, especially treebank based parsers, from performing well on such data. For this reason, we have developed the French Social Media Bank, the first user-generated content treebank for French, a morphologically rich language (MRL). The first release of this resource contains 1,700 sentences from various Web 2.0 sources, including data specifically chosen for their high noisiness. We describe here how we created this treebank and expose the methodology we used for fully annotating it. We also provide baseline POS tagging and statistical constituency parsing results, which are lower by far than usual results on edited texts. This highlights the high difficulty of automatically processing such noisy data in a MRL.

KEYWORDS: Treebanking, User Generated Content, Parsing, Social Media.

1 Introduction

Complaining about the lack of robustness of statistical parsers whenever they are applied on out-of-domain text has almost became an overused *cliché* over the last few years. It remains true that such parsers only perform well on texts that are comparable to their training corpus, especially in terms of genre. As noted by Foster (2010) and Foster et al. (2011b), most studies on out-of-domain statistical parsing have been focusing mainly on slightly different newspaper texts (Gildea, 2001; McClosky et al., 2006a,b), biomedical data (Lease and Charniak, 2005; McClosky and Charniak, 2008) or balanced corpora mixing different genres (Foster et al., 2007). The common point between these corpora is that they are edited texts. This means that their underlying syntax, spelling, tokenization and typography remain standard, even if they slightly depart from the newspaper genre. Therefore, standard NLP tools can be used on such corpora. Now, new forms of electronic communication have emerged in the last few years,namely social media and Web 2.0 communication media, either synchronous (micro-blogging) or asynchronous (forums), and the need for comprehensive ways of coping with the new languages types carried by those media is becoming of crucial importance.

In fact, the main consequence of the *Web 2.0 revolution* is that what was formerly restricted to one's inner circle of relations has now became widely available and is furthermore seen as containing potentially the same informativeness as written broadcast productions, that have undergone a full editorial chain, and that serve, most of the time, as the basis of our treebanks. Anyway, if those unlimited stream of texts were all written with the same level of proficiency as our canonical data source, the problem would be *simply*[1] a matter of domain adaptation. Yet, this is far from being the case 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, organized by Google, aimed at assessing the performances of parsers on various genres of Web texts (Petrov and McDonald, 2012).

Needless to say, such observations are likely to be even more true on web data written in morphologically rich languages (MRLs). These languages are already known to be arguably harder to parse than English for a variety of reasons (e.g., small treebank size, rich inflexion, free word order, etc.) exposed in details in (Tsarfaty et al., 2010). However, a lot of progress has been made in parsing MRLs using, for examples, techniques built on richer syntactic models, lexical data sparseness reduction or rich feature set. See (Tsarfaty and Sima'an, 2008; Versley and Rehbein, 2009; Candito and Crabbé, 2009; Green and Manning, 2010) to name but a few. The questions are thus to know: (1) to what extend MRL user generated content is *parsable*? and (2) more importantly, what is needed to fill that performance gap?

To answer question 1, we introduce the first release of the French Social Media Treebank, a representative gold standard treebank for French user-generated data. This treebank consists in around 1,700 sentences extracted from various types of French Web 2.0 user generated content (Facebook, Twitter, video games and medical board). This treebank was developed independently from the Google Web Treebank (Bies et al., 2012), the treebank used as development and test data for the above-mentioned SANCL shared task. In order to get first insights

[1] Cf. (McClosky et al., 2010) for numerous evidences of the non-triviality of that task.

into question 2, we provide a first set of POS tagging and parsing results using state-of-the-art systems trained on the French Treebank (FTB, Abeillé et al. (2003)), using our treebank as an evaluation corpus. These results show how difficult it is to process French user-generated data: for example, parsing results range from an astoundingly low 39.11% of labeled brackets F-score for the noisiest type of texts to 71-72% for better edited web parts — to be compared with the 86-89% regularly obtained on the FTB.

In the remaining of this paper, we first describe how we built the French Social Media Bank and the underlying motivations; we then introduce our annotation scheme which is based on the French Treebank (Abeillé et al., 2003) guidelines but extends it in many ways, due to the specificities of user-generated content. Next, we describe our annotation methodology, including the pre-processing tools that we developed and used, which were specifically adapted to deal with noisy texts. Finally, we discuss the results of baseline evaluation experiments on POS tagging, including results when using a dedicated wrapper for dealing with noisy texts, and constituency parsing. Since our tools were only trained on the FTB, which means that our results are baselines for future work based on our new French Social Media Bank.

2 Motivation and Corpus

As its English counterpart, the French web 2.0 generates a virtually unlimited stream of textual data. This term covers a wide range of practices, among which we decided to focus on microblogging (FACEBOOK and TWITTER) and on two types of web forums: one dedicated to general health issues for a wide public audience, DOCTISSIMO and one centered around video games issues (platform, help centers), JEUXVIDEOS.COM.[2] As we said in the introduction, we want to use these corpora to evaluate how difficult it is to parse raw user generated content in French and establish a realistic baseline using techniques we have successfully applied on well-written French texts.

To this end, we selected our corpora by direct examination through various search queries and ranked the texts according to our perception of how far they were from the French Treebank style (see below for details). We further added some very noisy texts to serve as a stress test for French statistical parsing. Table 1 presents some properties of our main corpora.

	# sent.	# tokens	avg. Length	std dev.
DOCTISSIMO	771	10834	14.05	10.28
high noisiness subcorpora	36	640	17.78	17.63
other subcorpora	735	10194	13.87	9.74
JEUXVIDEOS.COM	199	3058	15.37	14.44
TWITTER	216	2465	11.41	7.81
high noisiness subcorpora	93	1126	12.11	8.51
other subcorpora	123	1339	10.89	7.20
FACEBOOK	452	4200	9.29	8.17
high noisiness subcorpora	120	1012	8.43	7.12
other subcorpora	332	3188	9.60	8.49

Table 1: Corpus properties

Measuring noisiness 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

[2] `http://facebook.fr`, `https://twitter.com` (automatically configured to provide French tweets first based on IP address geo-localization), `http://forum.doctissimo.fr/` and `http://www.jeuxvideo.com/`

the French Treebank (Abeillé et al., 2003), we defined an ad-hoc *noisiness* metrics. It is simply defined as a variant of the Kullback–Leibler divergence[3] between the distribution of trigrams of characters in a given corpus and the distribution of trigrams of characters in a reference corpus.[4] As can be seen from Table 2, to which we added scores for the FTB dev and test sets for comparison purposes, our intuitions are confirmed by this metric. It shows that we cover various levels of noisiness, and that our reference corpus actually diverges more from the subcorpora that we have tagged as particularly *noisy*. As we shall see below, we have used this information for deciding for each subcorpus whether to pre-annotate it in a standard way or using a dedicated noise-tolerant architecture described in section in Section 5.1.

	noisiness score
DOCTISSIMO	0.37
high noisiness subcorpora	1.29
other subcorpora	0.31
JEUXVIDEOS.COM	0.81
FTB DEV	0.03
FTB TEST	0.003

	noisiness score
TWITTER	1.24
high noisiness subcorpora	1.46
other subcorpora	1.08
FACEBOOK	1.67
high noisiness subcorpora	2.44
other subcorpora	1.30

Table 2: Noisiness scores computed on tokenized version of the various sub-corpora. The (tokenized) FTB training set is used as a reference.

2.1 Corpus Overview

In this section we briefly introduce our corpora. All but the JEUXVIDEOS.COM corpus were collected in two phases: A first one dedicated to a light study of their contents (lexical differences, required level of preprocessing, etc.). At first glance, they seemed almost too edited and almost *too easy* for our parser. So in a second phase, we decided to look explicitly for texts harder to read and understand for average French speakers. We used French slang words in our search queries, including *verlan*[5] words, as well as urban youth idiomatic constructions such as *grave*[6] or *sa race*[7]. This lead to subcorpora that we found noisier, as was confirmed by the above-described metrics (see Table 2). In the case of the DOCTISSIMO part, we gathered texts from a forum dedicated to sexual intercourse problems between young adults. This choice was not without causing ethical concerns but given the fact that all private mentions were of course anonymized and all explicit references were filtered out, we ended up with 50 extremely noisy sentences, but greatly diverging from the newswire genre, and thus extremely

[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, Twitter hashtags and mentions 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," (iii) non-content sentences are ignored (e.g., tweet headers such as *Firstname Lastname @firstnamelastname*).

[4]Preliminary experiments on character bigrams and 4-grams have shown that the former are not informative enough and the latter lead to similar results than with trigrams. We also tried comparing distributions of tokens. Correlation with both intuition and parsing accuracy prove similar to that obtained with character 3-grams. However, token-based distribution divergences are less adequate for several reasons, among which: (i) correctly spelled unknown words affect more heavily token distributions than character-based distributions, whereas they should not affect the noisiness measure (e.g., they have a limited impact on tagging and parsing accuracy); (ii) character trigram distributions are less sparse than token distributions; (iii) there are more character trigrams than tokens in a sentence. Put together, the two last reasons show that a sound noisiness score on small corpora, or even at the sentence level, is more likely to be sound when working on character trigrams than on tokens.

[5]Very common French slang words where syllables are inverted to form new words which can in turn be *verlanized*. For instance, *arabe* 'arabic' is turned into *beur* which has been inverted again into *rebeu*.

[6]Post or pre-verbal intensifier adverb. *J'ai adoré grave!*, similar in meaning and style to *I totally enjoyed it!*.

[7]Post-verbal intensifier adverbial phrase, with the same usage as "one's a.. off."

interesting to evaluate our parsing chain.
Let us now describe and give examples for the various extracted subcorpora.

JEUXVIDEOS.COM [8] We collected data from 4 threads: *Call of Duty: Modern Warfare 3©*, *PC*, *Xbox 360©*, *Wii©* and *Linux*. Apart from spelling errors often involving phonetic spelling, which is found in all our corpora, this corpus is interesting because of the frequent use of English words, if not phrases, and a highly specialized lexicon.
In the examples below, the first line reproduces the original text, the second line is a standardized French version and the third line, an English translation attempt.

(1) a. Ces pas possible déjà que battelfield a un passe online
Ce n'est pas possible, Battlefield a déjà un pass en ligne
It's not possible, Battlefield already has an online pass

b. Si y'a que Juliet &Zayn qui sont co' sur le RPG,et qui font leur vie tranquilles
Si, il n'y a que Juliet et Zayn qui sont connectés sur le jeux de rôle et qui font leur vie tranquilles
Only Juliet and Zayn are connected on the RPG and are quiet doing their own business

DOCTISSIMO This corpus is made of two parts, each focusing on a different subtopic concerning birth control: patch birth control and pregnancy test related questions. These topics are populated by women of different ages and with different writing styles. The latter one being filled by younger women, the writing style is somewhat more sloppy. Moreover, as mentioned above, we added 50 extremely noisy sentences from the sexual intercourse section.

(2) a. pt que les choses ont changé depuis ?
Peut-être que les choses ont changé depuis ?
Maybe things have changed since then?

b. lol vu que 2-3 smaine apres qd jai su que j'etai enceinte jetai de 3 semaine.....
lol, vu que 2-3 semaines après, quand j'ai su que j'étais enceinte, je l'étais de 3 semaines....
Lol, given that 2-3 weeks later, when I learned I was pregnant, I was for 3 weeks...

c. car je ne me senté pa desiré, pa aimé, pa bel du cou, g t pa grd chose en fet.
Car je ne me sentais pas désirée, pas aimée, pas belle du coup, je n'étais pas grand chose en fait.
Because I didn't feel desired, nor loved, thus not beautiful, I wasn't much actually.

TWITTER This corpus is made of two parts: the first one focuses on news events of late November 2011.[9] Because all these tweets seemed to originate from semi-professional writers (mostly bloggers, journalists, politically engaged people), we built a second part with genuine non edited text. We used a list of keywords to gather such tweets and selected a balanced subset of those, as far as the style of writing is concerned.

(3) a. Je soupçonnes que "l'enfarineuse" était en faite une cocaineuse vu la pêche de #Hollande ce soir à #Rouen.
Je soupçonne que l'enfarineuse était en faite une cocaïneuse vu la pêche de #Hollande ce soir à #Rouen.
I suspect that the "flouring-lady" was actually a cocaïn-lady given the energy of #Hollande this night at #Rouen.

[8] Collected on November 9, 2011.

[9] An incident involving the left-wing candidate to the French presidency election, a so-called French hidden son of Adolf Hitler, and the then new right-wing election motto.

b. @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh
Ce n'est même pas élégant voyons, tu ne peux même pas marcher à coté de sa petite amie qu'ils font les voyous, ce n'est même pas leur rôle voyons.
It is not even elegant. One cannot even walk besides his girl friend, they already start bullying people. It is not even their role.[10]

FACEBOOK This corpus was built using publicly available comment threads on public profiles, with a focus on relatively known reality TV pseudo-artists known for their personal usage of French mixed with English. More texts were added using queries based on common first names (Sophie, Romain) and some French public personalities . One of the difficulties of FACEBOOK lies in the varying usage of the displayed login name in comments: it can either be part of the sentence (e.g., "[Spiderman] is tired") or not (e.g., "[Spiderman] I'm tired"). We decided to systematically keep these logins. We leave it to a post-processing step to remove them if appropriate. Note that the noisiest part of this corpus was not taken into account while adapting our POS tagger as described below.

(4) a. L' Ange Michael vraiment super conten pour toi mé tora plus grace a moi tkt love you !
L'Ange Michael: (Je suis) Vraiment super content pour mais tu auras plus grace à moi. Ne t'inquiètes pas. Je t'aime !
The Angel Michael: (I am) Really very happy for him but you'll get more because of me. Don't worry. I love you!

b. Afida: Viens on se check dans la vibes du moove pour voir comment on peut faire la hype à Hollywood avec Jane et Bryan
Afida: n/a
Afida: Come on, we'll check in into the moove's vibe to see how we can be hip in Hollywood with Jane and Bryan

3 Linguistics of user generated content

It is important to note that the aim of our work is to provide a sample of user-generated texts that are particularly difficult to parse for any parser based on an edited text treebank. It does not correspond to an single homogenous domain, although some specificities of user-generated content are found across various types of web data. Moreover, in some cases, and most notably TWITTER, such data include both linguistic content and media-specific meta-language. This meta-language (such as TWITTER's "RT" ("Retweet"), at-mentions and hashtags) is to be extracted before parsing *per se* or other types of linguistic processing. In this work, we focused on the linguistic content. Therefore, we deal with meta-language tokens only when they are embedded within or adjacent to purely linguistic content (e.g., the tweet itself, provided it consists of one or several sentences).

Prevalent idiosyncrasies in user generated content can be characterized on two axes: one which can be roughly describe as "the encoding simplification axis" which covers ergographic (1) and transverse phenomena (2) and the other "sentiment expression axis" which cover phenomena, qualified below as marks of expressiveness (3), emulating the same goal as sentiment expressed through prosody and gesture in direct interaction.

1. **Ergographic phenomena**, that is phenomenon aiming at reducing the writing effort, perceived as first glance as genuine misspell errors, cover in fact such a various set

[10]The translation is most certainly not accurate as even the original text is barely understandable for us.

of strategies it can be seen as a simplification of the encoding. Besides obvious typos, such as letter inversion[11], errors in letter doubling[12], wrong present participle[13] and so on, misspellings can be hard to categorize as such if they result in simpler word forms. This is why we include this category in the following list of phenomenon even if its intentionality is not always attested.

Phenomenon	Attested example	Standard counterpart	Gloss
a. Diacritic removal	*demain c'est l'***ete**	*demain c'est l'été*	'tomorrow this is summer'
b. Phonetization	*je suis* **oqp**	*je suis occupé*	'I'm busy'
c. Simplification	*je* **sé**	*je sais*	'I know'
d. Spelling errors	*tous mes* **examen son** *normaux*	*tous mes examens sont normaux*	'All my examinations are normal'

To this list we can also note the somewhat frequent omission of copula verbs and more generally different forms of elision (the subject pronoun, the negative adverb "ne"). Although not strictly lexical, this omission results also in less writing efforts. This is also noted in the Google Web Treebank where they compare this tendency in user generated content English to pro-drop languages' clitic elision.

2. **Transverse phenomenon: Contractions and typographic diaeresis.** Some phenomena can affect the number of tokens, compared to standard French, either by replacing several standard language tokens by only one, which we shall call a *contraction*, or conversely by splitting one standard language token into several tokens, called *typographic diaeresis*. Such phenomena are frequent in our corpora, and they need a specific annotation scheme (cf. Section 4). Note that the resulting non-standard tokens might be homographs of existing words, bringing more ambiguities if not properly analyzed.

   Contractions are way more diverse than in standard French (as instanced in the FTB). The only contractions that exist in standard French involve the prepositions *à* and *de* when followed by the definite article *le(s)* or the (rare) relative pronoun *le(s)quel(s)*. For instance, *à les* 'to the$_{\text{PLUR}}$' mandatorily becomes *au*.
   Within the FRENCH SOCIAL MEDIA BANK, we found sequences such as: (i) contraction between a subject clitic, a verbal form and a negation particle (e.g., *lapa* for *elle n'a pas*, 'she has not') ; (ii) contraction of interrogative verbal forms (e.g., *atu* for *as-tu*, 'have you') ; (iii) contraction and apocope of word compounds (e.g., *nimp* for *n'importe quoi*, 'rubbish'); (iv) contraction of determiners and nouns (e.g., *lesprit* for *l'esprit*, 'the spirit') and (v) contraction of object relative pronouns (or subordinate conjunction) and subject clitic (e.g., *qil* for *qu'il*, literally 'that he').

   Typographic diaeresis can be illustrated by *c a dire*, where these three tokens stand for the standard one-token conjunction *c'est-à-dire* (standing for "that is" or "namely"). It can also happen on top of a presumably already contracted token (e.g., *c t* for *ct/c'était*,'it was'). Note that many contractions and typographic diaeresis are built around a verb which is prefixed by a clitic. Therefore such contractions can project function labels and most of the time are the head of the sentence. Their mishandling can therefore propagate errors way beyond their immediate morpho-syntactic context and thus impacts parser performance very strongly.

[11] As in *J'ia* instead of *J'ai*/'I have'.
[12] e.g., *Dévelloppement* instead of *développement*/development
[13] e.g., "-ent" instead of "-ant"/ing, which are pronounced the same.

3. **Marks of expressiveness**
   Our treebank focuses in providing a sample of French Social Media web data, therefore most of its content describe dialogs and various forms of interaction between users through social media interface. Lacking ways of expressing sentiments, irony or anger through prosody and gesture, users deploy a wide range of strategy to add another dimension to their text stream. Most of them are evident, like graphical stretching, overuse of strong punctuation marks (as in "BEST. MOVIE. EVER." or "!!!!Greaaat!!!), abuse of emoticons, sometimes used as a verb (e.g., *Je t'<3* for *Je t'aime*,'I love you') or inside usernames, mixing between lowercase and uppercase and so on. Some are more anecdotal and tied to the particular type of software used to support a web forum, which allows the inclusion of url pointing to an animated picture, itself used to replace an emoticon. Needless to say that such urls add a considerable amount of noise in web forum texts. We list the main cases of such phenomenon in the table below.

Phenomenon	*Attested example*	*Standard counterpart*	Gloss
e. Punctuation transgression	*Joli !!!!!!*	*Joli !*	'nice!'
f. Graphemic stretching	*superrrrrrrrr*	*super*	'great'
g. Typographic transgression	*N* *U* *L*	*nul*	'bad'
h.Emoticons/smileys	*:-),<3*	–	–

Obviously the main effect of those different writing artifacts is to considerably increase the level of unknown words (compared to a treebank). More importantly, the new morphology brought by the those phenomenon complicates any processing based on regular unknown word identification through suffix analysis.
However, our general annotation strategy consists in staying as close as possible from the French Treebank guidelines (Abeillé et al., 2003) in order to have a data set as compatible, as much as possible, with existing resources.

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, syntagmatic 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 and added specific guidelines for handling idiosyncrasies user-generated content corpora.

4.1 FTB-UC vs. FTB

We targeted the annotation scheme of the FTB-UC (Candito and Crabbé, 2009), that was obtained by automatic modification of the FTB. The modifications with respect to the orignal FTB concern the tagset, the standardization of preposition and complementizer projections, and multi-word units:

- Multi-word units are very frequent in the FTB: 17% of the tokens belong to a compound. Compounds range from very frozen multi-word expressions like *y compris* 'including' (literally 'there included') to named entities. They include syntactically regular compounds with compositional semantics, such as *loi agraire* 'land law', that are encoded

as compounds because of a non-free lexical selection. These syntactically regular compounds tend to be inconsistently encoded in the FTB.[14] Further, the FTB includes "verbal compounds" that are potentially discontinuous, which provoke variable annotations. In the FTB-UC, these syntactically regular compounds are automatically mapped to a regular syntagmatic representation. We followed this rule for the annotation of the French Social Media Treebank. This has the virtue of uniformity, but clearly requires further treatment to spot clear cases of compounds (with non compositional semantics).

- Tagset: the tagset includes 28 POS tags —originally tuned by Crabbé and Candito (2008) to optimize parsing—, which are a combination of one of the 13 coarse-grained categories and other information that is encoded in the FTB as features, such as verbal mood information, proper versus common noun distinction, wh-feature, etc.
- Complementizers and prepositions: we annotate so that prepositions project a PP independently of the category of their object, contrary to the FTB's guidelines, in which prepositions with nominal objects project a PP, but those with infinitival objects don't. Further, we systematically use a sentential phrase as sister node to complementizers, contrary to the flat structure of the FTB.

Other notable additions to the annotation scheme concern the non-terminal tagset to which we added the FRAG label. It concerns phrases that cannot be syntactically attached to the main clause of a syntactic unit, e.g., mostly salutations, time stamp, meta sentential marks of emotion (emoticons, strong interjections). It also covers the case of usernames, at-mentions, and URL appended to a post/sentence (cf. FACEBOOK sentence 4a).

4.2 Additional extensions

A first extension needed for dealing with our data was to add two new POS tags, namely *HT* for TWITTER hashtags and *META* for meta-textual tokens, such as TWITTER's "RT". Note that TWITTER at-mentions as well as URLs and e-mail addresses have been tagged *NPP*. The rationale for this is to remain consistent with our tagging and parsing models trained on the FTB, which do not contain such tokens. This constitutes the main difference with other works on user-generated data.

However, the main extensions we added to the FTB annotation scheme are related to contraction and typographic diaeresis phenomena described in Section 3. The way we annotate (and automatically preannotate) such sequences is illustrated in Table 4. Let us now provide a few more details on each of these two cases.

Contracted tokens are associated with a combined POS tag which lists the sequence of each underlying words' tag. Let us illustrate this on *qil*, a non-standard contraction for *qu' il*. At least in some contexts, the tokens in the standard version *qu' il* would have been tagged respectively *CS* and *CLS*. In such contexts, the non-standard contracted token *qil* is tagged *CS+CLS*. Table 3 lists all such compound tags occurring more than twice in the treebank (37 more remaining).In some cases, such contractions involve two underlying forms, one being a verb and the other an argument of the verb (e.g., *jai* for *j' ai* 'I have'). In such cases, function labels are associated directly with the contracted token.

On the other hand, in cases of typographic diaeresis, the category of the multi-token standard counterpart is given to the last token, all others receive the special tag *Y*. Taking *c a dire* as an

[14] For instance *pays industrialisés* (*industrialized countries*) appears twice as a compound and 41 times as two words; *taux d'intérêt* (*interest rate*) appears 80 times as a compound and 25 times as two words.

Compound tag	Tag occ.	Attested example	Standard counterpart	Gloss
CLS+V	54	*c*	*c' est*	'it is'
ADV+CLO	12	*ni*	*n' y*	'(neg. adv.) (loc. clitic)'
CS+CLS	12	*qil*	*qu' il*	'that it/he'
CLS+CLO	11	*jen*	*j' en*	'I (gen. clitic)'
CLO+V	9	*ma*	*m' a*	'me$_{\text{dative}}$ has'
DET+NC	9	*lamour*	*l' amour*	'the love'
ADV+V	7	*non*	*n' ont*	'(neg. adv.) have$_{\text{3rd plur}}$

Table 3: Non-standard compound tags occurring at least 3 times.

example, which stands for the coordination conjunct *c'est-à-dire*, the first two tokens is tagged *Y* and *dire* is tagged *CC*. Note that this is consistent with the way such cases are annotated in the Google Web Treebank.[15]

Note that both phenomena can appear together. This is the case for example with *c t* instead of *c' était* 'it was' (tag sequence: *CLS V*): both letters *c* and *t* are used phonetically — as in using *U* for *you* in English — and sound like the two syllables of *c'é* and *tait*. Therefore, mapping *c* to *c'* and *t* to *était* is not adequate. In this case, we consider that a contraction followed by an typographic diaeresis has occurred, and associate *c* with the tag *Y* and *t* with *CLS+V*.

5 Annotation Methodology

We built manually validated treebank following a now well established methodology: we first defined a sequence of annotation layers, namely (i) sentence splitting, tokenization and POS tagging, (ii) syntagmatic parsing, (iii) functional annotation. Each layer is annotated by an automatic preprocessing that relies on previously annotated layers, followed by validation and correction by human annotators. At each step, annotators were able modify the choices made at previous stages. Our methodology is summarized as follows and detailed section 5.1:

- Segmentation, tokenization and POS tagging followed by manual validation and correction by one expert annotator.
- Constituency parsing followed by manual validation and correction by two annotators followed by an adjudication step.
- Functional annotation followed by manual validation and correction by two annotators followed by and adjudication step.

5.1 Pre-annotation strategies for the tokenization and POS layers

As mentioned above, we used two different strategies for tokenization and POS pre-annotation, depending on the noisiness score.

For less *noisy* corpora (those with a noisiness score below 1), we used a slightly extended version of the tokenization and sentence splitting tools from our standard FTB-based parsing architecture, Bonsai (Candito et al., 2010). This is because we want to have a tokenization that is as close as possible from the principles underlying the FTB's tokenization. Next, we used the POS-tagger MORFETTE (Chrupała et al., 2008) as a pre-annotator.

For corpora with a high noisiness score, we used a specifically developed pre-annotation process. This is because in such corpora, spelling errors are even more frequent, but also because

[15] In the Google Web Treebank, the counterpart of our tag *Y* is the tag *GW*.

the original tokens rarely match sound linguistic units, as can be seen on the example in Table 4 taken from the DOCTISSIMO file with the highest noisiness score. The idea underlying this pre-processing is to wrap the POS tagger (in this case, MElt) in such a way that it actually has to tag a sequence of tokens that is as close as possible to standard French, or, rather, to its training corpus (in this case, the FTB). Hence the following process, illustrated on a real example in Table 4:

1. We first apply several regular-expression-based grammars taken from the SxPipe pre-processing chain (Sagot and Boullier, 2008) for detecting smileys, URLs, e-mail addresses, Twitter hashtags and similar entities, in order to consider them as one token even if they contain whitespaces.
2. Next, we use the same tokenizer as for less *noisy* corpora.
3. We apply a set of 327 rewriting rules that were forged as follows: first, we extracted from our development corpus (all subcorpora but for the *noisy* Facebook subcorpus) *n*-gram sequences involving unknown tokens or occurring at an unexpectedly high frequency; then we manually selected the relevant ones and provided them manually with a corresponding "correction". The number of "corrected tokens" obtained by applying these rules might be different from the number of original tokens. In such cases, we use 1-to-*n* or *n*-to-1 mappings. For example, the rule *ni a pa* → *n'*␣*y a pas* explicitely states that *ni* is an amalgam for *n'* and *y*, whereas *pas* is the correction of *pa*.
4. We use the MElt tagger (Denis and Sagot, 2009), trained on the FTB-UC and the L*efff* lexicon (Sagot, 2010), for POS-tagging the sequence of corrected "tokens".
5. We apply a set of 15 generic and almost language-independent manually crafted rewriting rules, originally developed for English data (see below), that aim at assigning the correct POS to tokens that belong to categories not found in MElt's training corpus, i.e., the FTB; for example, all URLs and e-mail addresses are post-tagged as proper nouns whatever the tag provided by MElt; likewise, all smileys get the POS for interjections.
6. We assign POS tags to the original tokens based on the mappings between corrected POS-tagged tokens and original ones, and following the guidelines given in section 4.2. If a unique corrected token is mapped to more than one original tokens, all tokens but the last one are assigned the tag *Y*, and the last one receives the tag of the unique corrected token. If more than one corrected tokens are mapped to one original token, it is assigned a tag obtained by concatenating the tags of all corrected tokens, separated by the '+' sign. If the mapping is one-to-one, the POS tag provided by MElt for the corrected token is assigned to the corresponding original token.

This architecture is now available as part of the MElt distribution. It was also applied on English web data in the context of the SANCL shared task on parsing web data (Petrov and McDonald, 2012), with state-of-the-art results (Seddah et al., 2012).

5.2 Annotation strategy for constituency and functional annotation

Parse pre-annotation was achieved using a state-of-the-art statistical parser trained on the FTB-UC, provided with the manually validated tagging. The parser we used was the Berkeley parser (Petrov and Klein, 2007) adapted to French (Crabbé and Candito, 2008). Note that when the validated pos tags were discarded by the parser, in case of too many unknown word-pos pairs, those were reinserted.

To assess the quality of annotation, we calculated the inter annotator agreement using the

Original tokens	Gold corrected "tokens"	Automatically corrected and POS-tagged "tokens"	POS tags automatically assigned to the original tokens	Manually corrected POS tags for the original tokens
sa	ça	ça/*PRO*	sa/*PRO*	sa/*PRO*
fé	fait	fait/*V*	fé/*V*	fé/*V*
o moin	au_moins	au/*P+D* moins/*ADV*	o/*P+D* moin/*ADV*	o/*P+D* moin/*ADV*
6	6	6/*DET*	6/*DET*	6/*DET*
mois	mois	mois/*NC*	mois/*NC*	mois/*NC*
qe	que	que/*PROREL*	qe/*PROREL*	qe/*CS*
les	les	les/*DET*	les/*DET*	les/*DET*
preliminaires	préliminaires	preliminaires/*NC*	preliminaires/*NC*	preliminaires/*NC*
sont	sont	sont/*V*	sont/*V*	sont/*V*
sauté	sautés	sauté/*VPP*	sauté/*VPP*	sauté/*VPP*
c a dire	c'est-à-dire	c'est-à-dire/*CC*	c/*Y* a/*Y* dire/*CC*	c/*Y* a/*Y* dire/*CC*
qil	qu' il	qu'/*CS* il/*CLS*	qil/*CS+CLS*	qil/*CS+CLS*
yen	y en	y/*CLO* en/*CLO*	yen/*CLO+CLO*	yen/*CLO+CLO*
a	a	a/*V*	a/*V*	a/*V*
presk	presque	presque/*ADV*	presk/*ADV*	presk/*ADV*
pa	pas	pas/*ADV*	pa/*ADV*	pa/*ADV*

Table 4: Gold and automatic correction and POS tags for the following sentence extracted from the DOCTISSIMO file with the highest noisiness score 'Forplay have disappeared for at least 6 months, that is there is almost none.'

Parseval F-measure metric between two functionally annotated set of parses. Agreements range between 93.4 for FACEBOOK data and 97.44 for JEUXVIDEOS.COM (Table 5) and are on the same range than the DCU's Twitter corpus agreement score (Foster et al., 2011a). Similarly to that corpus, the disagreements involve fragments, interjections and the syntactic status to assign to meta-tokens elements. We note that our agreement scores are higher than those reported in other out-of-domain initiatives for French (Candito and Seddah, 2012). This small annotation error rate comes from the fact that the same team annotated both treebanks and was thus highly trained for that task. Maybe more importantly, social media sentences tend to be shorter than their edited counterparts so once POS tagging errors are solved, the annotation task is made relatively easier.

DOCTISSIMO	95.05	JEUXVIDEOS.COM	97.44
TWITTER	95.40	FACEBOOK	93.40
DCU'S TWITTERBANK	95.8	-	-

Table 5: Inter Annotator agreement

6 Preliminary experiments

Experimental Protocol In the following experiments, we used the FTB-UC as training data set, in its classical settings (test set: first 10%, dev set: next 10% and train set: the remaining.), see (Candito et al., 2009; Seddah et al., 2009) for details.

POS tagging experiments We have conducted preliminary evaluation experiments on the MElt POS-tagger(Denis and Sagot, 2009), used as such or within the normalization and correction wrapper described in the previous section. In Table 6, we provide POS-tagging accuracy results over the various subcorpora, following the DEV/TEST split described above. The results indicate that using the normalization and correction wrapper leads to significant improve-

ments in POS tagging accuracy. One can note that our accuracy results on standard TWITTER subcorpora are similar to the figures reported by Foster et al. (2011a) on English TWITTER data, although these figures are obviously not directly comparable, as they concern different languages using different tagsets. Another interesting observation is that these accuracy results are correlated with the noisiness metrics defined above.[16]

	DEV		TEST	
	MElt−corr	MElt+corr	MElt−corr	MElt+corr
DOCTISSIMO				
high noisiness subc.	56.41	80.78	–	–
other subcorpora	86.57	88.42	87.78	89.18
JEUXVIDEOS.COM	81.20	82.41	82.64	83.63
TWITTER				
high noisiness subc.	80.21	84.51	74.50	81.65
other subcorpora	84.09	89.00	86.23	88.24
FACEBOOK				
high noisiness subc.	–	–	67.00	70.75
other subcorpora	71.75	76.87	78.66	82.00
all	*80.64*	*84.72*	*83.10*	*85.28*
FTB (edited Text)	97.42	97.42	97.79	97.78

Table 6: Accuracy results for the MElt POS-tagger, embedded or not within the normalization and correction wrapper ("MElt+corr" and "MElt−corr" respectively). See text for details.

Baseline statistical parsing experiments In addition to the POS tagging experiments which showed that performance could greatly be improved using our normalization and correction wrapper, we performed a set of baseline experiments on the raw (tokenized) corpora using the PCFG-LA parser of Petrov et al. (2006) adapted to handle French morphology by (Crabbé and Candito, 2008). We used the PARSEVAL metrics applied to all sentences. Note that full scale experiments aimed at getting optimum parsing performance on this data set are out of the scope of this paper. We instead insist on providing baseline results, setting out a lower bound which assesses the difficulty of French User generated content parsing.
As expected, the results[17] provided in Table 7 show that there exists a large room for improvements. Interestingly, our user generated content data set seems even more difficult to parse than French biomedical data (67.79% vs 81.25% of F_1 score, on the Emea French test set for sentences of length lesser than 41), known to contain a high amount of unknown words and unusual phrase structures (Candito et al., 2011). Surprisingly, our parser performs poorly on FACEBOOK data, more than it does on TWITTER. In order to test their similarity, we can compare the respective noisiness scores of their subcorpora (FACEBOOK with respect to TWITTER = 1.42, TWITTER with respect to FACEBOOK 0.85). This shows that TWITTER data are more homogeneous than their FACEBOOK counterparts. Part of the reason lies in the inner nature of those social media: TWITTER is a live micro blogging platform, meaning that the content for a given trending topic shows fewer lexical divergences in a very short amount of time, whereas FACEBOOK public post are more distributed over time and posters.

The next step will involve collecting large unlabeled corpora to perform experiments with self-training techniques (McClosky and Charniak, 2008; Foster et al., 2011b) and unsupervised

[16]Simple linear regressions lead to the following results: without the normalization and correction wrapper, the slope is -4.8 and the correlation coefficient is 0.77; with the wrapper, the slope is -7.2 with a correlation coefficient as high as 0.88 (coefficients of determination are thus respectively 0.59 and 0.77).

[17]For convenience, we provide also baseline results on the FTB, see (Candito and Seddah, 2010).

	Dev set					Test set				
	LR	LP	F1	Pos acc.	OOVs	LR	LP	F1	Pos acc.	OOVs
Doctissimo										
high noisiness	37.22	41.20	39.11	51.72	40.47	-	-	-	-	-
other	69.68	70.19	69.94	77.96	15.56	70.10	71.68	70.88	79.14	15.42
JeuxVideos.com	66.56	66.46	66.51	74.56	20.46	70.59	71.44	71.02	75.70	19.88
Twitter										
high noisiness	62.07	64.14	63.09	64.89	31.50	54.67	58.16	56.36	64.40	32.84
other	68.06	69.21	68.63	79.70	24.70	71.29	73.45	72.35	78.88	24.47
Facebook										
high noisiness	-	-	-	-	-	55.26	59.23	57.18	54.64	50.40
other	55.90	58.71	57.27	64.34	38.25	60.98	61.79	61.38	70.68	29.52
all	64.13	65.48	64.80	72.69	23.40	66.69	68.50	67.58	74.43	22.81
Ftb	-	-	83.81	96.44	5.2	-	-	84.10	96.97	4.89

Table 7: Baseline parsing results split by sub corpora and noisiness level

word clustering within a PCFG-LA framework. Indeed, we have successfully applied these techniques for French out-of-domain parsing (Candito et al., 2011), as well as for parsing *noisy* English web data (Seddah et al., 2012). On the longer term we intend to apply our normalization and correction module before parsing. The parser will then be provided with corrected tokens, closely matching our regular training data, instead of unedited ones. This will compensate the lack of user generated content large unlabeled corpora, still lacking for French.

7 Conclusion

As mentioned earlier, the *French Social Media Bank* shares with the Google web bank a common will to extend the traditional treebank domain towards user generated content. Although of a smaller scale, it constitutes one of the very first usable resources to validate social media parsing and POS tagging, among the DCU Twitter and football BBC forums treebank (Foster et al., 2011a,b) and the Twitter data set from Gimpel et al. (2011). Moreover, it is the first set of syntactically annotated data for Facebook public web text.

Regarding the Google web bank, the way annotation guidelines had to be extended to deal with user generated content is largely consistent between both treebanks. However, our treebank differs from the Google Web Treebank in several aspects. First, French not only has a morphology richer than English, entailing a tedious disambiguation process when facing *noisy* data. Although the first version of our treebank is smaller than the Google Web Treebank, it includes richer annotations (compound POS, corrected token form of contractions) and includes subcorpora exhibiting a very high level of noise.

To conclude, we presented a new data set devoted on French user generated content. We proposed a first round of evaluation showing that simple techniques could be used to improve POS tagging performance. We presented baseline statistical parsing results, showing that performance on French user generated data were lying far behind those on newspaper in-domain texts. The take home message is that despite what is commonly said, parsing and POS tagging are far from being solved and that working on real text from real users is of crucial importance.

Acknowledgments We thank the reviewers for their insightful comments and Yoav Goldberg, Reut Tsarfaty and Jennifer Foster for their remarks on an earlier version of this work. This work was partly funded by the French ANR project EDyLex (ANR-09-CORD-008).

References

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

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English web treebank. Technical report, Linguistic Data Consortium,, Philadelphia, PA, USA.

Candito, M. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France. Association for Computational Linguistics.

Candito, M., Crabbé, B., and Seddah, D. (2009). On statistical parsing of french with supervised and semi-supervised strategies. In *EACL 2009 Workshop Grammatical inference for Computational Linguistics*, Athens, Greece.

Candito, M., Henestroza Anguiano, E., and Seddah, D. (2011). A word clustering approach to domain adaptation: Effective parsing of biomedical texts. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 37–42, Dublin, Ireland. Association for Computational Linguistics.

Candito, M., Nivre, J., Denis, P., and Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for french. In *Proceedings of COLING 2010*, Beijing, China.

Candito, M. and Seddah, D. (2010). Parsing word clusters. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

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.

Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco. ELDA/ELRA.

Crabbé, B. and Candito, M. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, pages 45–54, Avignon, France.

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.

Elsner, M. and Charniak, E. (2011). Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1179–1189. Association for Computational Linguistics.

Foster, J. (2010). "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and van Genabith, J. (2011a). # hardtoparse: Pos tagging and parsing the twitterverse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Nivre, J., Hogan, D., and van Genabith, J. (2011b). From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 893–901, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Foster, J., Wagner, J., Seddah, D., and Van Genabith, J. (2007). Adapting wsj-trained parsers to the british national corpus using in-domain self-training. In *Proceedings of the Tenth IWPT*, pages 33–35.

Gildea, D. (2001). Corpus variation and parser performance. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202, Pittsburgh, USA.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, DTIC Document.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Green, S. and Manning, C. (2010). Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402. Association for Computational Linguistics.

Lease, M. and Charniak, E. (2005). Parsing biomedical literature. *Natural Language Processing–IJCNLP 2005*, pages 58–69.

McClosky, D. and Charniak, E. (2008). Self-training for biomedical parsing. In *Proceedings of ACL-08: HLT, Short Papers*, pages 101–104, Columbus, Ohio. Association for Computational Linguistics.

McClosky, D., Charniak, E., and Johnson, M. (2006a). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.

McClosky, D., Charniak, E., and Johnson, M. (2006b). Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics.

McClosky, D., Charniak, E., and Johnson, M. (2010). Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Petrov, S. and McDonald, R. (2012). Overview of the 2012 Shared Task on Parsing the Web. In *Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), a NAACL-HLT 2012 workshop*, Montreal, Canada.

Sagot, B. (2010). The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of LREC'10*, Valetta, Malta.

Sagot, B. and Boullier, P. (2008). SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2).

Seddah, D., Candito, M., and Crabbé, B. (2009). Cross parser evaluation and tagset variation: A French Treebank study. In *Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09)*, pages 150–161, Paris, France. Association for Computational Linguistics.

Seddah, D., Sagot, B., and Candito, M. (2012). The Alpage Architecture at the SANCL 2012 Shared Task: Robust Preprocessing and Lexical bridging for user-generated content parsing. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, Montréal, Canada.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., and Tounsi, L. (2010). Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12, Los Angeles, CA, USA. Association for Computational Linguistics.

Tsarfaty, R. and Sima'an, K. (2008). Relational-realizational parsing. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 889–896. Association for Computational Linguistics.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for german. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 134–137, Paris, France. Association for Computational Linguistics.