Spanish Treebank Annotation of Informal Non-standard Web Text

Mariona Taulé[2(✉)], M. Antonia Martí[2], Ann Bies[1], Montserrat Nofre[2], Aina Garí[2], Zhiyi Song[1], Stephanie Strassel[1], and Joe Ellis[1]

[1] Linguistic Data Consortium, University of Pennsylvania, 3600 Market Street, Suite 801, Philadelphia, PA 19104, USA
[2] CLiC, University of Barcelona, Gran Via 588, 08007 Barcelona, Spain
mtaule@ub.edu

Abstract. This paper presents the Latin American Spanish Discussion Forum Treebank (LAS-DisFo). This corpus consists of 50,291 words and 2,846 sentences that are part-of-speech tagged, lemmatized and syntactically annotated with constituents and functions. We describe how it was built and the methodology followed for its annotation, the annotation scheme and criteria applied for dealing with the most problematic phenomena commonly encountered in this kind of informal unedited web text. This is the first available Latin American Spanish corpus of non-standard language that has been morphologically and syntactically annotated. It is a valuable linguistic resource that can be used for the training and evaluation of parsers and PoS taggers.

1 Introduction

In this article we present the problems found and the solutions adopted in the process of the tokenization, part-of-speech (PoS) tagging and syntactic annotation of the Latin American Spanish Discussion Forum Treebank (LAS-DisFo).[1] This corpus consists of a compilation of textual posts and includes suggestions, ideas, opinions and questions on several topics including politics and technology.

Like chats, tweets, blogs and SMS these texts constitute a new genre that is characterized by an informal, non-standard style of writing, which shares many features with spoken colloquial communication: the writing is spontaneous, performed quickly and usually unedited. At the same time, to recover the lack of

This material is based on research sponsored by Air Force Research Laboratory and Defense Advanced Research Projects Agency under agreement number FA8750-13-2-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government.

[1] A Discussion Forum is an online asynchronous discussion board where people can hold conversations in the form of posted messages.

F. Daniel and O. Diaz (Eds.): ICWE 2015 Workshops, LNCS 9396, pp. 15–27, 2015.
DOI: 10.1007/978-3-319-24800-4_2

face-to-face interactions, the texts contain pragmatic information about mood and feelings often expressed by paratextual clues: emoticons, capital letters and non-conventional spacing, among others. As a consequence, the texts produced contain many misspellings and typographic errors, a relaxation of standard rules of writing (i.e. the use of punctuation marks) and an unconventional use of graphic devices such as the use of capital letters and the repetition of some characters.

These kinds of texts are pervasive in Internet data and pose difficult challenges to Natural Language Processing (NLP) tools and applications, which are usually developed for standard and formal written language. At the same time, they constitute a rich source of information for linguistic analysis, being samples of real data from which we can acquire linguistic knowledge about how languages are used in new communication modalities. Consequently, there is an increasing interest in the analysis of informal written texts, with annotated corpora where these characteristics are explicitly tagged and recovered as one of the crucial sources of information to fill this need. In particular, this Latin American Spanish Treebank is being developed in support of DARPA's Deep Exploration and Filtering of Text (DEFT) program, which will develop automated systems to process text information and enable the understanding of connections in text that might not be readily apparent to humans. The Linguistic Data Consortium (LDC) supports the DEFT Program by collecting, creating and annotating a variety of informal data sources in multiple languages to support Smart Filtering, Relational Analysis and Anomaly Analysis.

This paper is structured as follows. After a brief introduction to the related work (section 2), we present how the LAS-DisFo was built (section 3). Then, we describe the annotation process carried out (section 4), followed by the annotation scheme and criteria adopted (section 5). First, we focus on the word-level tokenization and morphological annotation (subsection 5.1) and, then, on the sentence segmentation (subsection 5.2) and syntactic annotation (subsection 5.3). Final remarks are presented in (section 6).

2 Related Work

It is well known that NLP tools trained on well-edited texts perform badly when applied to unedited web texts [7]. One of the reasons for this difficulty is the result of a mismatch between the training data, which is typically the Wall Street Journal portion of the PennTreeBank [11] in the case of English, and the corpus to be parsed. Experiments carried out with English texts such as those reported in [13] show that current parsers achieve an accuracy of 90% when they are limited to heavily edited domains, but when applied to unedited texts their performance falls to 80%, and even PoS tagging scores only slightly higher than 90%. The problem increases with morphologically rich languages such as French [14] and Spanish.

Considering that many NLP applications such as Machine Translation, Sentiment Analysis and Information Extraction need to handle unedited texts, there

is a need for new linguistic resources such as annotated web text corpora to extend already existing parsers and for the development of new tools.

The annotation of unedited web corpora presents specific challenges, which are not covered by current annotations schemes and require specific tagsets and annotation criteria. This explains the increasing interest in the organization of workshops focusing on the annotation of informal written texts (EWBTL-2014; NLPIT-2015; LAW-Informal text-2015). There is an increasing interest in the development of annotated corpora of non-standard texts. These are usually small corpora in which the different web genres are represented or representative of one specific genre: English Web Treebank [2]; French Social Media Bank [14]; the No-Sta-D corpus of German non-standard varieties [6]; the #hardtoparse corpus of tweets [8], among others.

3 Latin American Spanish Discussion Forum Corpus

3.1 LDC Spanish DF Data Collection

Spanish discussion forum (DF) data was collected by LDC in support of the DEFT program, in order to build a corpus of informal written Spanish data that could also be annotated for a variety of tasks related to DEFT's goal of deep natural language understanding. DF threads were collected based on the results of manual data scouting by native Spanish speakers who searched the web for Spanish DF discussions according to the desired criteria, focusing on DF topics related to current events and other dynamic events. The Spanish data scouts were instructed to search for content on these topics that was interactive, informal, original (i.e., written by the post's author rather than quoted from another source), and in Spanish (with a particular focus on Latin American Spanish during the latter part of the collection). After locating an appropriate thread, scouts then submitted the URL and some simple judgments about the thread to a collection database via a web browser plugin. Discussion forums containing the manually collected threads were selected and the full forum sites were automatically harvested, using the infrastructure described in [9].

3.2 Latin American Spanish DF Data Selection and Segmentation

A subset of the collected Spanish DF data was selected by LDC for annotation, focusing on the portion that had been harvested from sites identified as containing primarily Latin American Spanish. The goal was to select a data set suitable for multiple levels of annotation, such as Treebank and Entities, Relations, and Events (ERE) [15]. Creating multiple annotations on the same data will facilitate experimentation with machine learning methods that jointly manipulate the multiple levels. Documents were selected for annotation based on the density of events, which was required for ERE. The resulting Latin American Spanish DF data set to be used for Spanish Treebank annotation consists of 50,291 words and 2,846 sentences in 60 files, each of them a thematically coherent fragment from a forum.

4 Annotation Process

The LAS-DisFo corpus is annotated with morphological and syntactic information by applying automatic and manually annotation processes. Firstly, the corpus was automatically tokenized, PoS tagged and lemmatized using tools from the Freeling library[2] [12]. Then, a manual check of the output of these automatic processes was carried out. At this level, a greater level of human intervention was required than with standard written corpora. As we will observe in the annotation criteria sections, most of the problems arose from word tokenization and word spellings rather than at the syntactic level.

LAS-DisFo was then subjected to a completely manual syntactic annotation process. In order to guarantee the quality of the results, we first carried out the constituent annotation followed by the annotation of syntactic functions.

The annotation team was made up of seven people: two senior researchers with in-depth experience in corpus annotation that supervised the whole process; one senior annotator with considerable experience in this field, who was responsible for checking and approving the whole annotation task; and four undergraduate students in their final year, who carried out the annotation task. One of the students reviewed the morphology, two students annotated constituents and the other two students annotated both constituents and functions. This organization meant that the earlier annotations were revised at every stage of the process. After one and a half months of training, the three syntactic annotators carried out an interannotator agreement test using 10 files. These files were manually compared and we discussed solutions for the inconsistencies that were found, so as to minimize them. The initial guidelines were updated and the annotation process started. The team met once a week to discuss the problems arising during the annotation process to resolve doubts and specific cases.

The annotations were performed using the AnCoraPipe annotation tool [1] to facilitate the task of the annotators and to minimize the errors in the annotation process. The corpora texts annotated were XML documents with UTF-8 encoding.

5 Annotation Scheme and Criteria

Two main principles guided the whole annotation process. First, the source text was maintained intact. The preservation of the original text is crucial, because in this way the corpus will be a resource for deriving new tools for the analysis of informal Spanish language, as well as for the linguistic analysis of spontaneous written language. Second, we used a slightly modified version of the annotation scheme followed for the morphological and syntactic tagging of the Spanish AnCora corpus ([3]; [16]) and we extended the corresponding guidelines ([2]; [10]) in order to cover the specific phenomena of non-standard web texts. In this way, we ensure the consistency and compatibility of the different Spanish resources.

[2] http://nlp.lsi.upc.edu/freeling/

The main differences in the annotation scheme are due to the addition of special paratextual and paralinguistic tags for identifying and classifying the different types of phenomena occurring in this type of texts (misspellings, emphasis, repetitions, abbreviations, and punctuation transgressions, among others) and the criteria to be applied for dealing with them. However, the AnCora tagset has not been modified with new morphological or syntactic tags.

A summary of the criteria applied in the annotation of LAS-DisFo is presented below. We describe the criteria followed for word-level tokenization and its corresponding PoS tagging and then those applied for sentence-level tokenization and syntactic annotation.

5.1 Word-Level Tokenization

Most of the problems in the annotation process arose from word tokenization and word spellings. Therefore, the tokenization and morphological annotation processes required considerable effort. The kind of revision carried out consisted of addressing problems with word segmentation, verifying and assigning the correct PoS and lemma to each token, and resolving multiword expressions. The PoS annotation system[3] is based on [3].

Below, we present the criteria adopted in order to resolve the phenomena encountered in the discussion forum texts, which we have organized in the following groups: 1) word-segmentation phenomena; 2) typos and misspellings; 3) abbreviations; 4) marks of expressiveness; 5) foreign words, and 6) web items.

1. **Word-Segmentation Phenomena.** This kind of error mostly results from speed writing errors. As a general criterion, we always preserve the word form of the source text, except when the spelling error involves two different words with an incorrect segmentation, when the two words appear joined (1) or when a word is wrongly split due to the presence of a blank space (2). In these cases, the original text is modified. We justify this decision because this was a rare phenomenon, with an anecdotic presence in the corpus, and correcting these errors allowed for the correct PoS and syntactic annotation. In examples of word games,[4] we respect the original source and treat them like a multiword expression (if the words are split).

   (1) Esto **estan** de incrédulos. (instead of **es tan**)
   'This **isso** like incredulous people ...' (instead of **is so**)
   word=es lemma=ser pos=vaip3s
   word=tan lemma=tan pos=rg

   (2) Sistema de **gener ación** de bitcoins (instead of **generación**)
   'System for **gener ating** bitcoins' (instead of **generating**)
   word=generación lemma=generación pos=ncfs

[3] http://www.cs.upc.edu/~nlp/tools/parole-eng.html
[4] In the data annotated no word games were found.

In example (1) the criterion applied is to split the incorrect segment into two words, whereas in example (2) the criterion is to join the two segments into one word. In both cases, we assign the corresponding lemma and PoS tag to each word.

2. **Typos and Misspelling Errors.** The main typos found involve the omission/insertion of a letter, the transposition of two letters (3), the replacement of one letter for ano-ther, wrongly written capital letters, and proper nouns or any word that should be in capital letters but that appears in lower case. We also treat as typos those involving punctuation marks, usually a missing period in ellipsis (4).

   (3) **pre**sonas '**pre**sons' (instead of **per**sona, '**per**son')
   word=presonas lemma=persona pos=ncfp000 anomaly=yes

   (4) pero**..** lo bueno 'but**..** the best thing' (instead of **...**)
   word=.. lemma= ... pos=fs anomaly=yes

   In the case of misspellings, the most frequent mistakes are related to diacritic/accent removal, which normally also results in an incorrect PoS tag (5), but the omission of the silent 'h' in the initial position of the word, or the use of 'b' instead of 'v' (or vice versa), corresponding to the same phoneme, are also frequent. Dialectal variants (6), which are not accepted by the Royal Spanish Academy of Language, are also considered misspellings.

   (5) todo cambi**o**...'all change' (instead of todo cambi**ó** 'everything changed')
   word=cambio lemma=cambiar pos=vmis3s anomaly=yes

   (6) **a**moto (instead of moto, 'motorbike')
   word=amoto lemma=moto pos=ncfs anomaly=yes

   In example (5) the omission of the diacritic involves the assignment of an incorrect PoS, both 'cambio' and 'cambió' are possible words in Spanish, the former is a noun and the latter a verb, therefore the analyzer tagged 'cambió' as a noun. In this case, we manually assigned the correct verbal PoS (vmis3s) and the corresponding verbal lemma (infinitive, cambiar 'to change'), without modifying the original form.

   The criteria adopted to resolve these phenomena is to maintain the source text, assign the correct PoS and lemma and add the label 'anomaly=yes' for typos and misspellings. In this way, the different written variants of the same word can be recovered through the lemma, and the typos and misspelling words are also easily identified by the corresponding labels.

3. **Abbreviations.** This kind of phenomena results in a simplification of the text aiming at reducing the writing effort. The abbreviations encountered

usually involve the omission of vowels, but consonants can also be omitted (7). In these cases, we assign the correct PoS and lemma and add the label 'toreviewcomment=a[5]' for identifying them.

(7) **tb** me gusta escribir 'I also like to write' (**tb** instead of también)
forma=tb lemma=también pos=rg toreviewcomment=a

4. **Marks of Expressiveness.** One of the phenomena that characterizes informal non-standard web texts is the unconventional use of graphic devices such as emoticons (8), capital letters (9) and (10), and the repetition of characters (11) to compensate for the lack of expressiveness in the writing mode. These are strategies that allow us to get closer to the direct interaction of oral communication. We use different labels and criteria to annotate the different types of marks of expressiveness:

   For emoticons, we assign the lemma describing the emoticon with the prefix 'e-' and the PoS 'word', which indicates unknown elements.

   (8) **:)**
   word=:) lemma=e-contento ('e-happy') pos=word

   For words in capital letters indicating emphasis (9) and for the emphatic repetition of vowels and other characters within words (10), we add the label 'polarity_modifier=increment'. We also assign the label 'toreviewcomment=cl[6]', when a fragment or an entire paragraph is written in capital letters (11). In this case, we add the label at the highest node (phrase or sentence).

   (9) es algo totalmente **NUEVO**! 'is something totally **NEW**!'
   word=NUEVO lemma=nuevo pos=aqms polarity_modifier=increment

   (10) **muuuuy** grande**!!!** '**veeeery** big**!!!**' (instead of muy grande!)
   word=muuuuy lemma=muy pos=rg polarity_modifier=increment
   word=!!! lemma=! pos=fat polarity_modifier=increment

   (11) LOS OTROS, LOS Q NO APORTAN, NO SE GANARÁN NI UN SEGUNDO D MI TIEMPO Y MI ESCRITURA.
   'THE OTHERS, WHO DO NOT CONTRIBUTE ANYTHING, WILL NOT HAVE A SECOND OF MY TIME OR MY WRITING'.

   (LOS OTROS, LOS Q NO APORTAN, NO SE GANARÁN NI UN SEGUNDO D MI TIEMPO Y MI ESCRITURA.) <sentence toreviewcomment=cl polarity_modifier=increment>

[5] 'a' stands for abbreviation.
[6] 'cl' stands for capital letters.

5. **Foreign Words.** In this kind of text the presence of words (12) or fragments written in another language (13), usually in English (and especially in technical jargon), is frequent. The criterion followed in these cases is not to translate the words to Spanish, and we add the label 'wdlng[7]=other'. In the case of fragments, we assign a simplified PoS tag (just the category) and all the words are grouped in a fragment at a top node (sentence, clause (S) or phrase).

   (12) Estás **crazy**? 'Are you **loco**?' ('crazy' instead of loco)
   word=crazy lemma=crazy pos=aqcs000 wdlng=other

   (13) you are my brother
   word=you lemma=you pos=p wdlng=other
   word=are lemma=are pos=v wdlng=other
   word=my lemma=my pos=t wdlng=other
   word=brother lemma=brother pos=n wdlng=other

   Syntactic annotation: (you are my brother)<sentence>

6. **Web Items**. We include website addresses, URLs, at-signs before usernames, and other special symbols used in web texts such as hashtags[8] in this category. Following the same criteria used in the AnCora annotation scheme, we tagged these web items as proper nouns and named entities with the value 'other'.

   (14) http://www.afsca.gob.ar
   word=http://www.afsca.gob.ar lemma= http://www.afsca.gob.ar pos=np ne=other

5.2 Sentence Segmentation

The LAS-DisFo corpus was automatically sentence segmented in the PoS tagging process by the CLiC team, and the resulting segments were then manually corrected. It is worth noting that this level of segmentation required considerable human intervention because in informal web texts the use of punctuation marks frequently does not follow conventional rules: we found texts without any punctuation marks; texts that only used 'commas' as marks; texts with an overuse of strong punctuation marks usually for emphatic purposes, and texts with wrongly applied punctuation marks. These non-conventional uses lead to the erroneous automatic segmentation of sentences. Therefore, before starting with syntactic annotation it is necessary to correct the output of the automatic segmentation. The criteria followed are described hereafter.

[7] 'wdlng' stands for word language.
[8] In the data annotated no hashtags were found.

1. We apply normal sentence segmentation (15) when final punctuation (period, question mark, exclamation mark, or ellipsis) is correctly used. When ellipsis is used as non-final punctuation (16), we do not split the text.

   (15) (Hubieron dos detenidos por robos en medio del funeral...)<sentence>
   '(Two people were arrested for robberies in the middle of the funeral...')<sentence>

   (16) (Las necesidades no las crearon ellos solos... tambien ayudo el embargo)<sentence>
   '(The needs did not create themselves... it also helped the embargo')<sentence>

2. We do not split the text into separate sentences when final punctuation marks (usually periods) are wrongly used (17). If periods are used instead of colons, commas, or semicolons, we consider the text to be a sentence unit and we add the label 'anomaly=yes' to the punctuation mark.

   (17) (Los cambios que debería hacer Capitanich. Integrar publicidad privada. Cambiar a Araujo.)<sentence verbless=yes>
   '(The changes that Capitanich should make. Integrate private advertising. Switch to Araujo.)<sentence verbless=yes>'

   In example (17), the first period should be a colon and the second period should be a semicolon or a coordinated conjunction. In both cases, they are tokenized and tagged as periods (PoS=fp[9]) with the label 'anomaly=yes'. This sentence unit is treated as a <verbless> sentence because the main verb is missing.

   When the emoticons (18) are at the end of the sentence, they are included in the same sentence unit.

   (18) (Ni idea :?( )<sentence>
   '(No idea :?( )'<sentence>

3. We split the text into separate sentences when final punctuation marks are not included (19) and when a middle punctuation mark is used instead of final punctuation marks (20). In the former case, we add an elliptic node (∅) with the labels 'pos=fp', 'elliptic=yes' and 'anomaly=yes'. In the latter case, the label 'anomaly=yes' is added to the erroneous punctuation mark.

   (19) (Lo bueno debe prevalecer ∅<name=fp> <elliptic=yes> <anomaly=yes>)
   '(Good must prevail ∅<name=fp> <elliptic=yes> <anomaly=yes>))'

[9] 'fp' stands for punctuation period.

(20) hoy ya no pueden hacerlo, la tecnologia los mantiene a rayas**,,**
(hoy ya no pueden hacerlo, la tecnologia los mantiene a rayas**,** <PoS=fc> <anomaly=yes> **,** <PoS=fc> <anomaly=yes>)<sentence>
'(today, they can no longer do so, the technology keeps them in line**,** <PoS=fc> <anomaly=yes> **,** <PoS=fc> <anomaly=yes>)<sentence>

In example (20), the second comma could be interpreted either as an ellipsis or as a repeated period. The context of this sentence points to the second interpretation.

In addition to the commas incorrectly used as final punctuation marks, many other problems appear in the sentence. In the example above the first word of the sentence appears in lowercase instead of uppercase, the accent is missing in 'tecnologia' and 'rayas' should be written in singular (See section 5.1).

5.3 Syntactic Annotation

Regarding syntactic annotation, we followed the same criteria that we applied to the AnCora corpus [16], following the basic assumptions described in [4]: the annotation scheme used is theory-neutral; the surface word order is maintained and only elliptical subjects are recovered; we did not make any distinction between arguments and adjuncts, so that the node containing the subject, that containing the verb and those containing verb complements and adjuncts are sister nodes.

We adopted a constituent annotation scheme because it is richer than dependency annotation (since it contains different descriptive levels) and, if it is necessary, it is easier to obtain the dependency structure from the constituent structure. Syntactic heads can be easily obtained from the constituent structure and intermidiate levels can be avoided [5].

It was agreed to tag only those syntactic functions corresponding to sentence structure constituents, whether finite or non-finite: only subject and verbal complements were taken into consideration. We defined a total number of 11 function tags, most of them corresponding to traditional syntactic functions: subject, direct object, indirect object, prepositional object, adjunct, agent complement, predicative complement, attribute, sentence adjunct, textual element and verbal modifier.

When it was necessary to syntactically annotate more than one sentence within a sentence unit (for instance, embedded clauses like relative, completive and adverbial clauses), they were included under the top node <sentence>. In the same way, embedded sentences were tagged as <S> with the feature <clausetype> instantiated, its possible values being <completive>, <relative>, <adverbial> and <participle>.

The syntactic annotation of LAS-DisFo did not present as large a variety of phenomena as the morphological annotation did, but we did find many differences with respect to formal edited texts. In the discussion forum texts, the

Table 1. Syntactic information in LAS-DisFo and IS-NW

Corpus	Words	Sentences	Verbless	Discontinuities	Inserted elements
IS-NW	50,988	2,049	281	59	44
LAS-DisFo	50,291	2,846	1,229	139	98

frequency of verbless sentences, incomplete sentences, discontinuities and parenthetical elements (that did not belong to the general structure of the sentence) is higher than in news-based corpora such as IS-NewsWire[10]. Fragments without a main verb are treated as verbless sentences. In the case of LAS-DisFo, it is worth noting that these verbless sentences can be the result of joining several fragments of texts separated by wrongly used punctuation marks (17). Table 1 shows a comparision of these phenomena in the LAS-DisFo and LAS-NW corpora.

6 Final Remarks

In this paper, we have presented the criteria and annotation scheme followed in the morphological and syntactic annotation of the LAS-DisFo corpus, which contains 50,291 words and 2,846 sentences. Discussion Forum texts, like other kind of web texts, are characterized by an informal, non-standard style of writing. This results in texts with many misspellings and typographic errors and with a relaxation of the standard rules of writing. Furthermore, they usually contain pragmatic information about mood and feelings, often expressed by paratextual clues. All these characteristics pose difficult challenges to NLP tools and applications, which are designed for standard and formal written language.

The main challenges in the annotation of these kinds of texts appear in the segmentation of lexical and syntactic units and in the treatment of all the variants found at word level. To our knowledge, this is the first morphologically and syntactically annotated corpus of Spanish informal texts. This corpus will be released through the LDC catalog, and will be a new resource that could prove useful for deriving new tools for the analysis of informal Spanish language and Latin American Spanish, as well as for the linguistic analysis of spontaneous written language.

[10] The International Spanish Newswire TreeBank (IS-NW) consists of 50,988 words selected from the Spanish Gigaword previously released in LDC2011T12. The IS-NewsWire corpus has been also annotated with syntactic contituents and functions following the AnCora guidelines by the same annotator team. IS-NW and LAS-DisFo constitute the LDC Spanish Treebank, including the first Latin American Spanish corpus with morphological and syntactic annotation.

References

1. Bertran, M., Borrega, O., Martí, M.A., Taulé, M.: AnCoraPipe: A new tool for corpora annotation. Tech. rep., Working paper 1: TEXT-MESS 2.0 (Text-Knowledge 2.0) (2010). http://clic.ub.edu/files/AnCoraPipe_0.pdf
2. Bies, A., Mott, J., Warner, C., Kulick, S.: English Web Treebank. Linguistic Data Consortium, Philadelphia (2012)
3. Civit, M.: Criterios de etiquetación y desambiguación morfosintáctica de corpus en español. Coleccion de monografias de la SEPLN (2003)
4. Civit, M., Martí, M.A.: Design principles for a Spanish treebank. In: Proceedings of Treebanks and Linguistic Theories (2002)
5. Civit, M., Martí, M.A., Bufí, N.: Cat3LB and Cast3LB: from constituents to dependencies. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 141–152. Springer, Heidelberg (2006)
6. Dipper, S., Lüdeling, A., Reznicek, M.: NoSta-D: A corpus of German Non-Standard varieties. Non-standard DataSources in Corpusbased Research. Shaker Verlag (2013)
7. Foster, J.: "cba to check the spellig" investigating parser performance on discussion forum post. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, pp. 381–384 (2010)
8. Foster, J., Çetinoglu, Ö., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., Van Genabith, J.: # hardtoparse: POS tagging and parsing the twitterverse. In: AAAI 2011 Workshop on Analyzing Microtext, pp. 20–25 (2011)
9. Garland, J., Strassel, S., Ismael, S., Song, Z., Lee, H.: Linguistic resources for genre-independent language technologies: user-generated content in BOLT. In: Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey (2012)
10. Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., Eskander, R.: Developing an Egyptian Arabic treebank: impact of dialectal morphology on annotation and tool development. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland (2014)
11. Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The penn treebank: annotating predicate argument structure. In: Proceedings of the Human Language Technology Workshop, San Francisco (1994)
12. Padró, L., Stanilovsky, E.: FreeLing 3.0: towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey, May 2012
13. Petrov, S., McDonald, R.: Overview of the 2012 shared task on parsing the web. In: Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), vol. 59. Citeseer (2012)
14. Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V.: The French social media bank: a treebank of noisy user generated content. In: COLING 2012–24th International Conference on Computational Linguistics, Mumbai, pp. 2441–2458 (2012)

15. Song, Z., Bies, A., Riese, T., Mott, J., Wright, J., Kulick, S., Ryant, N., Strassel, S., Ma, X.: From light to rich ERE: annotation of entities, relations, and events. In: Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015), Denver (2015)
16. Soriano, B., Borrega, O., Taulé, M., Martí, M.A.: Guidelines: Constituents and syntactic functions. Tech. rep., Working paper: 3LB (2008). http://clic.ub.edu/corpus/webfm_send/17