ISSN 2007-9737

Social Media – Processing Romanian Chat and Discourse Analysis

Cătălina Mărănduc[1, 2], Cenel-Augusto Perez[1], Radu Simionescu[1]

[1] Faculty of Computer Science, Al. I. Cuza University, Iași, Romania

[2] Academic Institute of Linguistics Iorgu Iordan – Al. Rosetti, Bucharest, Romania

{catalina.maranduc, augusto.perez, radu.simionescu }@info.uaic.ro

Abstract. In order to obtain a balanced corpus, a sub-corpus of 2,576 sentences illustrating contemporary social media language has been added to the Dependency Treebank for Romanian. The texts were taken from the chat. The subject of this paper is to describe the second step of processing non-standard texts with a hybrid POS-tagger for Romanian and with a Malt parser, both until now trained on standard language and on other styles of communication. The results obtained show that the UAIC tools are comparable with the tools for other languages trained on similar corpora. Another purpose is to develop this resource, the Dependency Treebank for Romanian, not only quantitatively, doubling its dimension in a year, but also changing its format with a new one, compatible with other similar foreign corpora, and adding new, more complex annotation layers. A semantic layer and a discursive annotation will be added, permitting the study of discursive and conversational particularities. Finally, examples illustrating discursive particularities of the chat communication are discussed.

Keywords. Conversational particularities, dependency treebank, discourse analysis, processing non-standard texts, social-media communication.

1 Introduction

Romanian Language is yet difficult to process, because it has an insufficient number of representative resources. There are few resources, either small or insufficient, or inconsistently annotated, or inaccessible to researchers. A new resource is easy to begin, but difficult to increase without inconsistencies; it is also very difficult to maintain. The format must be changed to one compatible with universal similar corpora. Adding new levels of more complex annotation is difficult, but necessary in order to maintain the researcher's interest.

The UAIC Romanian Dependency treebank has now 12,885 sentences and 239,550 words (punctuation included). It has a complex automatic morpho-syntactic annotation, entirely, carefully, manually checked. A program for the automatic transposition in UD (Universal Dependencies) format has been built at the Research Institute of Artificial Intelligence, with the financing of a grant and the participation of our Faculty of Computer Science[1], and the texts transposed are being supervised by linguists.

In fact, the objective was included in the protocol of the grant, but we do not know if it has been fulfilled. We do not have access to such a transposition program (UAIC to UD conventions), and will probably have to build one. One of the authors of the present article has built a program for changing the format of the documents from XML to CONLL-U and vice versa. Most of the old UAIC-Ro treebank, built before December 2014, was, however, included in the "Romanian Treebank" affiliated at UD, together with roughly equal numbers of sentences, annotated at RACAI (Research Institute of Artificial Intelligence in Bucharest) in UD conventions.

The purpose of the project cited in note 1, is to build a syntactic parser, trained in the Universal Dependencies system of annotation and with a

[1]"Semantics-driven Syntactic Parser for Romanian" (SSPR).

better accuracy than the UAIC Malt parser variant. Our goal is different, namely to increase the UAIC Malt parser performance by increasing the number of sentences in the corpus gold for the training, and to include in the gold corpus as many non-standard versions of the Romanian. The UAIC conventions of annotation contain a greater number of syntactic-semantic information.

The affiliation of UD allows conducting comparative studies with similar resources for the other 30 languages that participate to the UD. This system of annotation is based on the Pennsylvania Treebank annotation. It pays attention especially to the mandatory verbal dependencies, which form the core of the clause. Words and clauses with the same syntactic relation are differently annotated. An important purpose of the UD project is to build a language independent syntactic parser. However, the UAIC-RoDepTb[2] has a more complex annotation, and developing a corpus should not mean reducing the amount of information annotated and renouncing of some features mandatory for rendering Romanian language-specific phenomena.

Examples:

– The Romanian language has a lot of reflexive pronouns with different meanings, characteristic for combinatorial valences of verbs and defining for the verb's meaning: *a se uita* "to look at" and *a uita* "to forget" are different meanings disambiguated only by the reflexive pronoun. The UD conventions have no relation for annotating the reflexive, situations like the above being inexistent in English. The reflexive pronoun was annotated with the label "expl:pv", which means that it resumes the information about the subject that the verbal predicate contains, so the reflexive repeat an unexpressed information. We believe that the chosen solution is absurd, but we will comply, as it is a convention to be consistently followed.
– Prepositions also are determinant for the combinatorial valences of verbs; intransitive verbs cannot appear without a specific preposition; consequently, the prepositional object is mandatory, but in UD it is considered a facultative dependency (modifier). In the UD conventions adopted for Romanian, the label nmod: pmod was chosen for the prepositional dependencies which are mandatory in Romanian. They are opposed to a large number of nmod with preposition (non-mandatory) that are not nmod:pmod (mandatory, although "mod" is abbreviation for modifier, a non- mandatory dependency). However, ultimately, all the conventions are questionable.
– In Romanian there are five or six different types of weak pronouns that repeat the information of another word, noun or pronoun; more of them can appear in the same clause; consequently, a sub-classification of expletives would be required for correctly annotating the co-references, but the UD system does not offer solutions for these situations. An "expletive impersonal" (accepted by UD) does not exist in Romanian. In our language, the impersonal is expressed by a reflexive pronoun in the accusative case without a nominative subject.
– Example: *Se pare că*… (It seems that…).

The information annotated and supervised for all the 12,885 sentences will be kept in other layers of annotation, parallel with the layer transposed in UD conventions, in the same way that the PDT (Prague Dependency Treebank) is organized.

UAIC-RoDepTb is a balanced corpus, illustrating all the styles of the Romanian language. The treebank has been increased with a lot of quotations from the big Romanian Thesaurus Dictionary[3], in diverse styles of the language, but obsolete (from the eighteenth and nineteenth centuries). To illustrate the contemporary language, a solution was to add a sub-corpus of social-media communication, annotating conversation from the chat. It is a new style of communication with a rapid increase and with similitudes with the oral style. The texts chosen are written by people with a higher education. The specifics of this kind of communication is the non-standard, free, creative expression of ideas, and

[2] http://nlptools.infoiasi.ro/resources/

[3] eDTLR = the project no. 1696 / 2007-2010, led by Dan Cristea, transposed in electronic format the Academic Thesaurus Dictionary (http://85.122.23.96/).

the tools, especially the POS-tagger, are trained on standardized texts but the non-standard text processing is made with difficulties.

2 Related Work

Social media communication has been thoroughly studied, with various purposes, for sentiment analysis, for information retrieval, for counting the occurrences of neologisms, and so on. However, the POS-tagging (the classification and analysis of parts of speech) is an indispensable step for any research. Social media communication is similar with oral communication, having many lexical inventions. The authors of the papers on this subject describe methods to increase the accuracy of POS-taggers on non-standardized texts.

In the papers [11, 112], the researchers describe a new corpus for German, composed of 36,000 annotated tokens, which contain web comments. The dimensions of the corpus are comparable with the sub-corpus for Romanian described below. In these papers a Marcov model POS-tagger is presented. Another POS-tagger trained on non-standardized texts is the syntactic-semantic Bayesian HMM, described in [5]. The POS-taggers are trained on a particular language, for example German, French, English, Chinese, Italian, Indonesian, etc.

The work in [134] aims to formulate a model to develop a POS-tagger using the clustering of the words contained in the corpus for automatically improving the quality of machine learning. Many papers, [1, 4, 9, 134], describe Twitter texts processing systems. In [13], the Frog tagger is combined with a post-processing module that incorporates in the lexicon the new Twitter-specific tags found in the Frog part-of-speech output. Approximately 1 million tweets collected in the context of the SoNaR project were tagged by Frog and the post-processor combined.

There are two procedures to increase the accuracy of social media processing tools: the inclusion of new modules and rules in the tools for the non-standardized text processing like in [18], or the "normalization" of the social media texts [6], [10], transforming non-standardized text into standardized ones. However, in this way, the particularities of this kind of communication are removed, and not studied. Other procedures experienced in adapting a POS-tagger for noisy texts are described in the papers [8, 9, 17].

The number of research papers that study the social media parsing ([4, 7]) is smaller than the number of those studying the POS-tagging. Parsers can analyze a sentence's grammatical structure, being influenced by the previously POS-tagger annotations. The authors of [4] wrote hundreds of rules to account for hash tags, repeated letters, and other linguistic features specific of Twitter communication. Their program classified correctly 68% of 2,304 tweets.

3 Corpus and Tools Presentation

The social-media sub-corpus has been processed and supervised in two steps, resulting in documents with the dimensions shown in the Table 1. The first 200 sentences on 09_chat1.xml have been automatically annotated, then manually checked, by typing for each word the lemma in the standard language, this being considered as the gold corpus; subsequently, the supervised chat_1 was added to the gold corpus, by the bootstrapping method of training.

The first step of the training of the natural language processing tools on the chat has been presented in a communication of the RUMOUR Workshop[4] [15]. Then, the 10_chat_2.xml, automatically annotated and manually checked, was also added to the training corpus. Finally, a little corpus for discourse experiments, having 131

Table 1. Quantitative presentation of the Chat sub-corpus

Title of document	Number of sentences	Number of words	Average words per sentence
09_chat_1	944	13,243	14.02
10_chat_2	1,500	24,278	16.18
chat_final	131	1,773	13.53
Total	2,575	39,294	15.25

[4] http://eurolan.info.uaic.ro/2015/events/workshop/

Table 2. First evaluation of UAIC Romanian POS tagger for the social media texts (2015)

Diacritics only, eval. on 1984	Diacritics only, eval. on 200 sent. from the corpus	Mixed diacritics, eval. on mixed 1984	Mixed diacritics, eval. on the first 200 sentences from the chat corpus
97.03 %	54.33%	94.38%	74.53%

sentences and 1,773 tokens, was processed with very good results; it was called chat_final.

– Beside the occasional inventions of the free communication, in Romanian chat there are letters with diacritics (ă, î, â, ș, ț), generally omitted, but not always. A clone of the UAIC hybrid POS-tagger[5] [16] was trained on the *Nineteen Eighty-Four* Orwell's novel (translation in Romanian), a gold corpus created in the MULTEXT-East[6] project, doubling the corpus, without Romanian diacritics. The result of experiments is shown in Table 2.

The training corpus with and without diacritics contains more new ambiguities in comparison with the standard language, which explains the results above.

The corpus is processed also with the UAIC parser[7], a variant of the Malt parser [13], trained on standard texts, and then on both standard and nonstandard texts, with the results shown in the Table 3.

All the sub-corpora from the UAIC-RoDep Treebank and the social media corpus 09_chat_1.xml have been brought together in a 10-fold scheme. The results presented in Table 3 clearly demonstrate the importance of the training on social media non-standardized texts.

These results contain the precision of the head detection (column 4), the precision of the label (dependency relation) attachment (column 5) and the precision of both these parameters (column 3)

In the second step of the processing, the documents 09_chat_1, 10_chat_2 and chat_final have been entirely manually checked by experts. The word forms and the lemmas have been introduced in the lexicon of the POS-tagger. The introduction of rules for the ambiguities created is yet in course.

We obtained the results in Table 4:

The Malt parser for Romanian is trained now on a gold corpus of 12,885 sentences, including the chat_1-2 sub-corpora and evaluated on the chat_final, all sentences being annotated and supervised observing the UAIC conventions of annotation. The parser has now an increased accuracy, and it will not be abandoned after the creation of a parser to annotate Romanian sentences in UD conventions. Increasing the training corpus and including in it a bigger corpus of social media communication, we obtained increased parser accuracy, as shown in Table 5.

The corpus consistently annotated is automatically transposed in the UD conventions of annotation. The transposition results are being supervised by linguists, in order to create a big gold corpus for the training of the new parser.

4 Analysis of Social Media Discourse Peculiarities

In the paper [15] some comments about the stylistic and pragmatic peculiarities of Social Media have been analyzed: regarding the lexical contamination of words with ironic and playful intention, for example, or observing in which modality the communicative roles are established (by banal or creative modalities). In this paper, some discourse characteristics will be analyzed.

[5] UAIC Romanian POS-tagger: http://nlptools.infoiasi.ro/WebBinPosRo.

[6] http://nl.iis.si/ME.

[7] UAIC Romanian dependency parser http://nlptools.infoiasi.ro/WebFdgRo/.

Table 3. Evaluation of UAIC Romanian syntactic parser (2015)

Metrics	Both attachment	Head attachment	Label attachment
Trained on standard texts	Standard texts	78.04%	84.25%	83.57%
	Social media	58.31%	71.74%	66.08%
Trained on standard and nonstandard texts	Standard texts	77%	83.65%	82.99%
	Social media	66.01%	78.48%	72.21%

Table 4. Evaluation of the mixed diacritics POS-tagger trained on social media

Evaluation of 09_chat_1.xml	Evaluation of 10_chat_2.xml	Evaluation of chat_final
83.33%	87.12%	89.43%

Table 5. Evaluation of the UAIC parser on social media

Metrics	Both attachment	Head attachment	Label attachment
Trained on 4,800 standard sent. eval.on Chat_1	66.01%	78.48%	72.21%
Trained on 9,420 sent. Chat_1, included eval. on Chat_2	69.33%	80.23%	75.55%
Trained on 12,885 sent. including Chat_1&2, eval. on Chat_final	89.11%	93.67%	91.76%

A particularity of this style is caused by the written form. While one person is typing a reply, he or she cannot read what his/her interlocutor wrote, perhaps he/she opens another theme of conversation, and later he/she receives the answer of his/her previous reply. Often the conversation is about several interwoven themes.

The intertwining of themes is a well-studied phenomenon in chat conversations. See, for instance, the paper "The Right Frontier Constraint Holds Unconditionally" by Dan Cristea [3]. This paper describes the rule of the new structures, always or habitually attached only in the right part of the tree.

In the sub-corpus chat_1-2, consecutive sequences of replies have been annotated, so that it is possible to select more sentences with consecutive ids for making a discourse or a conversational analysis.

Example 1:

(B) sentence id. 805: Dar de ce? Nu vrea frumușelul să îl văd? (*But why? The fancy boy does not want me to see him?*)

(C) sentence id. 806: Azi am scos pisica afară de două ori. (*Today I took the cat out twice*)

(B) sentence id. 807: E la un coleg, socializează. (*He's at a colleague's, he is socializing*)

(C) sentence id. 808: Era foarte speriată, s-a dus sub smochin. (*She was very scared; she went under the fig tree*)

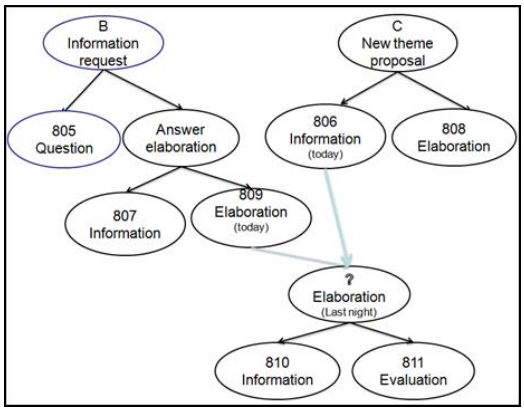

Fig. 1. The discourse tree of the sentences 805–811

(B) sentence id. 809: A venit acasă, și-a lăsat ghiozdanul și a plecat. (*He came home, he left his backpack and left*)

(?) sentence id. 810: Aseară tot așa a făcut. (*Last night he/it? did the same thing*).

(?) sentence id. 811: Aha, cred că se va obișnui în timp. (*Ah yes, I think he/it? will get used to it in time*)

The replies 805, 807, 809 refer to a boy, and the replies 806, 808 refer to a cat. The last two refer to the cat, but they are ambiguous, because in Romanian the subject is elliptical. If we relate these replies in another way, we obtain some absurd funny stories:

806–807 = the cat socializes with a colleague;

808–809 = the cat left his backpack and went under the fig tree;

809–811 = the boy gets used over time to leave home every night.

It is possible that the interlocutors may observe these false and inadequate links and amuse themselves. The interlocutor will add the reply:

(B) sentence id. 812: Daa? Ai accepta ca el să facă asta în fiecare noapte? (*Ah, yes? Would you accept his doing this every night?*)

In order to find a tool for analyzing this kind of communication, we propose the veins theory, presented by Dan Cristea for a discourse model in the chapter "Motivations and Implications of the Veins Theory" [2].

We must specify that these examples are neither a discourse, nor a text. The discourse is the text presented in a communication situation. The length of the string is not important. The text/discourse is defined by a global sense and by a global intention of communication. If a person is in the street and cries: "The car!!!" addressing the communication to another person crossing the street, in terms of pragmatics, this is a discourse because it has a global sense and a clear communicative intention, well understood by the receiver.

Another condition for considering a string as a text or a discourse is that it should have a saturated informational degree, especially of its limits, at the start and at the end. In the above fragment of conversation, the information is insufficient for understanding the two global senses: 1. "the boy is drunken (and he cannot pay a visit)" and 2. "the cat does not go outside to defecate." Of course, the sense is clear for the two communicators, because they have some information after the conversation.

Therefore, we can study the discursive relations of these syntactically independent sentences, which do not constitute a discourse as such, by building the discourse trees.

The semantic incongruence between 806–807, and between 808–809 is evident. Therefore in Romanian the subject is elliptical, the cat has neither a colleague nor a back-pack, this relation being comical and absurd. These are two unrelated trees. However, the third sub-tree can be connected either to the tree B, or to the tree C. The sentences 806 and 809 render events that happened on the day of the dialog. The subject is elliptical. The adverbial modifier "the same thing" refers to an event reported before, either in the tree B, or in the tree C. The single mark for the interpretation is the sentiment analysis for the second interlocutor.

The word "socializes" is ironically selected, because it is correctly used in the professional communication, and not in familiar contexts. This is a mark of the lack of agreement of the second communicator with the action of the boy. However, the last reply: "Ah yes, I think..." is an approval of the information in the sentence 810, so it is positively evaluated. The action of "gets used over time to leave home every night" is positive for the cat and negative for the boy; therefore, we can attach the third tree to the second tree.

In agreement with the veins theory, the interpretation of the third sub-tree as a continuation

of the first one is less likely, because this theme is related to peripheral information; the first theme has as its root the question about the causes of the refusal of the boy to pay a visit (to be seen). This question has no direct answer. The head of the second level of the structure is the euphemistic elaboration of the motivation, without the direct affirmation that he is drunk; the elaboration contains an ironical selection of the word "socialize" and some unnecessary details about his back-pack. It will be a mistake in the composition of the discourse to attach a new unnecessary elaboration by generalizing the disagreeable situation.

On the contrary, the second theme has as its root the cat's exit from the house, the elaboration contains details about its reactions and the second level of the tree related the repetition of the action in the first sentence. This repetition has the purpose to accustom the cat with the action (desirable), and the last reply is related with the root and the head of the second level.

In the example above, the conversation is difficult to follow, without making replicas belonging to the first or second participant. Another kind of annotation, a discursive one, is experienced on the little sub-corpus chat_final, with the intention to extend it to the other sub-corpora chat 1-2 and to other fragments in the treebank, also formed by related sentences.

Example 2:

1.<SPEAKER1>**Nik:**<OPENING>Bonsoaree şi piper ! (*Bon soiree* (Fr) *and pepper*!)<OPENING/>

<QUESTION1>Spune-mi, rogu-te, cine iaşte una, Lorenţa Nuştiucum[1], de a scris o carte despre vocabularul modei ? (*Tell me, please, who is a certain Lorente Dontknowhow[1], who wrote a book about the fashion vocabulary*?) <QUESTION1/><SPEAKER1/>

2.<SPEAKER2>**Ugla:**<OPENING>Una bună și ție! (*One good for you, too!*)<OPENING/>

<ANSWERquest1>E pretena[1] cu noi, e nevasta[1] lui Laponu, adica Victor popescu, un țicnit fost coleg de liceu cu nevasta ta. (*She[1]'s our friend, the wife[1] of the Lapon, alias Victor popescu, a crazy man who went to high school with your wife*.) <ANSWERquest1/>

<TOPIC1a>E lucrare de doctorat. (*It's a doctoral dissertation*)

<TOPIC1a/><SPEAKER2/>

3. <SPEAKER1>**Nik:**<TOPIC1b>Se vede. (*It could understand it*.) <TOPIC1b/><SPEAKER1/>

4<SPEAKER2>**Ugla:**<ANSWERquest2>Cond usa de Mareana Neț[2]. (*Supervisor Marianne Neț[2]*.)<ANSWERquest2/><SPEAKER2/>

5<SPEAKER1>**Nik:**<QUESTION2>la SalaM ? (*SalaM supevisor*?) <QUESTION2/>

<TOPIC1c>Ai răspuns. (*You answered.*) <TOPIC1c/>

<TOPIC2a>Aia[2] mi-amintesc cum că ierea o vita brevis. (*I remember that she[2] was a vita (life ? cow ?) brevis* [Lat: *short*]). <TOPIC2a/><SPEAKER1/>

6<SPEAKER2>**Ugla:**<CONFIRMATION1c>Da . (*Yes.*) <CONFIRMATION1c/>

<TOPIC1d>Ea[1] a lucrat bebelotecara la Bebeloteca Academiei au pensionat-o[1] si ci-că o[1] reangajeaza. (*She[1] worked as a librarian at the Romanian Academy and retired, but they say that she[1] was to be re-employed*.) <TOPIC1d/><SPEAKER2/>

7. <SPEAKER1>**Nik:**<QUESTION3topic2a> Mareana[2]? (*Marianne[2]?*)<QUESTION3topic2a/>

<QUESTION4topic1d>Pă cine ? (*who?*) <QUESTION4topic1d/><SPEAKER1/>

8. <SPEAKER2>**Ugla:** <ANSWERquest3>E de treaba, de la cercu de poetica (*She[2]'s amiable, from the circle of poetics*) <ANSWERquest3/><SPEAKER2/>

9. <SPEAKER1>**Nik:**<QUESTION4repet> pă cine angajează? (*Who do they re-employ?*) <QUESTION4repet/>

<CONFIRMTIONtopic2c> Da. (*Yes*) <CONFIRMTIONtopic2c/><SPEAKER1/>

10. <SPEAKER2>**Ugla:**<TOPIC2d>A publicat in strainatati. (*She[2] published abroad*.) <TOPIC2d/>

<TOPIC1e>Cre ca au si angajat-o[1] pt ca am vazt carti de la ea[1] pe birou la sefa. (*I think they have employed her[1] because I saw books brought by her[1] on the desk of my boss*.) <TOPIC1e/><SPEAKER2/>

11. <SPEAKER1>**Nik:**<QUESTION5top2e>Pă Neţ[2]? (*They employed Neţ[2]?*) <QUESTION5top2d/> <SPEAKER1/>

12. <SPEAKER2>**Ugla:**<TOPIC1f>Ea[1] imprumuta carti care ne trebe noua si ni le da sa le scanam, (*She[1] borrows books we need and gives them to us to scan*.) <TOPIC1f/><ANSWERquest4>Lorența[1]. (*Lorence[1]*) <ANSWERquest4/>

<TOPIC2e>Neț[2] nu are treaba, e CS I. (*Neț[2] has no problems, she[2] is CS I*.) <TOPIC2e/><SPEAKER2/>

13. <SPEAKER1>**Nik:**<TOPIC1f>Aha, la pensii ? (*Aha, to pensions*?) <TOPIC1f/><SPEAKER1/>

In this example there are two speakers that practice a lot of creative untranslatable archaic, popular, familiar, bookish expressions. The playful intention is obvious.

We experienced a conversational and discursive .xml annotation, marking the following categories: <SPEAKER>, <QUESTION><ANSWER><CONFIRMATION>, <INFIRMATION>, <OPENING>, <CLOSSING>, <TOPIC> and adding them numbers or other indications to establish the relations between the categories. The fragment cited does not contain the closure of the replies exchanged. These annotations will be applied on the syntactic layer or on the semantic layer, on a previously annotated text, but in this paper the annotation is simplified for space saving reasons.

In this example there are two topics. The first speaker is interested in a book author (TOPIC1) and after the answer that introduces the person to him, another person is named, the PhD supervisor of the first person (TOPIC2).

For easier reading, in this paper the first person was marked with number 1 superscript, and the second person with number 2 superscript; also the 13 replies were numbered, so we can refer to them, and the topic 2 was aligned within the page. The two speakers have extra-text information to disambiguate the two topics, such as: they can refer to the person with the title CS I also identified as a doctoral supervisor retired at a more senior age; also, they have the possibility of asking questions to clarify the appurtenance of a new information at the topics.

However, a computer will have more difficulties to find the co-references in this text. Perhaps a way to solve this task should be to consider the appearance of she^1 in contexts with *our friend, we need, give us* (see 2, 12) and she^2 in contexts with *abroad*, *circle of poetics*, not familiar for the speakers (see 8, 10).

While speaker 1 is typing the question about the supervisor, the answer to that question is already posted, preceding it, that being also a specific feature of the chat communication (see 4-5).

Further, in the same reply, the second speaker gives information about topic1 and information about topic2, maybe with the playful intention of confusing the first speaker, or because the new question about TOPIC2 appears while the information about TOPIC1 is typed (see 10, 12).

The first speaker is more interested in TOPIC2, therefore he introduced TOPIC1, and the second speaker accumulates a lot of information not required about TOPIC1. The two speakers seem not to understand each other; therefore there exists a similitude in their enjoyment of playing with words. The ambiguity of the communication is adjusted by more questions (see 7, 9).

Nik seems to accept the playful intention of Ugla to hide the identity of persons who she offers unsolicited amalgamated information. He asks questions about the retirement of Neț, although he knows that for a supervisor it is at a later age. The answer to the first speaker's questions comes at the end of the conversation, resembling the deciphering of a riddle, and Nik explains his satisfaction and comprehension by citing a famous humorist's reply.

5 Conclusions and Future Work

The Social Media is an important means of communication in contemporary society, with a fast increasing volume on the web and with specific peculiarities. That is why these texts must be included in the corpora for each natural language. In [3, 4] the authors describe a new corpus for German, composed of 36,000 annotated tokens, which contain web comments. Generally, these corpora are formed of a single kind of Social Media communication, either the blogs, the chat, the comments, or the twitter. A bigger corpus with more types of SM should be interesting, for the comparative study of peculiarities of each of them. However, the automatic annotation of this kind of texts is very difficult, and the supervision of the result is time consuming.

The tools for natural language processing are built for standardized communication, and the Social Media is the non-standardized empire of the absolute freedom of communication. Supporting only the pragmatic rules, that message is constrained only by the need to optimize the

transmission to the recipient of the content (often affective) intended by the transmitter. The transmitter uses the most creative stylistic means to achieve the communicative purpose.

The standard rules of the language, built to facilitate the communication, accepted by all the speakers, start to trivialize to such an extent in use, that they cannot express the communicative and precise intentions anymore; they are suspended by the communicators endowed with high creativity.

Since the sub-corpus is more different from that which our tools were trained on, we need a larger training corpus in the social media style, in order to operate satisfactorily and to increase the accuracy of the results.

The results obtained in the second step of development of the corpus are better, and after the training of the parser with about 2,500 chat sentences, the actual results of the parsing on chat_final represent a qualitative jump. The POS-tagger has higher difficulties than the syntactic parser in processing the chat; the creative examples must be introduced in its lexicon and more rules for the disambiguation must be written. The discursive annotation system proposed below is minimal; we intend to refine and apply the discursive annotation on the semantic layer of annotation that will be implemented on the entire corpus UAIC-RoDepTb.

References

1. **Avontuur, T., Balemans, I., Elshof, L., van Noord, N., & van Zaanen, M. (2012).** Developing a Part-of-Speech Tagger for Dutch Tweets. *Computational Linguistics in the Netherlands Journal,* Vol. 2, pp. 34–51.
2. **Cristea, D. (2005).** Motivations and Implications of Veins Theory. **Bernadette Sharp (Ed.)***. Natural Language Understanding and Cognitive Science, Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*, NLUCS, in conjunction with ICEIS 2005, Miami, U.S.A., INSTICC Press, Portugal, pp. 32–44.
3. **Cristea, D. (2005).** The Right Frontier Constraint Holds Unconditionally. *Proceedings of the Multidisciplinary Approaches to Discourse (MAD'05),* Chorin/Berlin, Germany.
4. **Dent, K., Alto, P., & Diep, F. (2011).** Parsing the Twitter verse, *Scientific American* 305, 22
5. **Darling, W., Paul, M., & Song, F. (2012).** Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1–9.
6. **Derczynski, L., Maynard, D., Aswani, N., & Bontcheva, K. (2013).** Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, pp. 21–30. DOI: 10.1145/2481492.2481495.
7. **Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., & van Genabith, J. (2011).** POS Tagging and Parsing the Twitter verse. *Proceedings of the AAAI Workshop on Analyzing Microtext*.
8. **Gadde, P., Subramaniam, L., & Faruquie, T. (2011).** Adapting a wsj Trained Part-of-Speech Tagger to Noisy Text: Preliminary Results. *Proceedings of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data.* ACM. DOI: 10.1145/2034617.2034623.
9. **Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. (2011).** Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 42–47. ACL.
10. **Liu, F., Weng, F., & Jiang, X. (2012).** A Broad Coverage Normalization System for Social Media Language. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1035–1044.
11. **Neunerdt, M., Trevisan, B., Reyer, M., & Mathar, R. (2013).** Part of Speech Tagging for Social Media Texts. **Gurevich, I., Biemann, C., & Zesch, T. (Eds):** GSCL 2013, LNAI 8105, pp. 139–150. Springer Verlag Berlin Heidelberg. DOI: 10.1007/978-3-642-40722-2_15.
12. **Neunerdt, M., Reyer, M., & Mathar, R. (2013).** A POS Tagger for Social Media Texts trained on Web Comments. *Polibits*, Vol. 48, pp. 59–66.
13. **Nivre, J., Hall, J., & Nilsson, J. (2006).** MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, May 24-26, 2006, Genoa, Italy, pp. 2216–2219.
14. **Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., & Schneider, N. (2012).** Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances.

Technical Report CMU-ML-12-107, Machine Learning Department, Carnegie Mellon University.

15. **Perez, C.A., Mărănduc, C., & Simionescu, R. (2016).** Including Social Media – a Very Dynamic Style, in the Corpora for Processing Romanian Language. **Trandabăţ, D. & Gîfu, D. (Eds.):** *Proceedings at EUROLAN 2015*, CCIS 588, Springer International Publishing Switzerland, pp. 139–153. DOI: 10.1007/978-3-319-32942-0_10.
16. **Simionescu, R. (2011).** Hybrid POS Tagger. *The Workshop on Language Resources and Tools in Industrial Applications*, Eurolan.
17. **Singha, K.R., Purkayastha, B.S., & Singha, K.D., (2012).** Part of Speech Tagging in Manipuri: A Rule-based Approach. *International Journal of Computer Applications,* Vol. 51, No.14.
18. **Toutanova, D., Klein, C., Manning, C., & Singer, Y. (2003).** Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 173–180, ACL. DOI: 10.3115/1073445.1073478.

Cătălina Mărănduc recieved his first PhD in general linguistics from The University of Bucharest, Romania. Now she is PHD student at the Faculty of Computer Science at the Al. I. Cuza University of Iași, Romania.

Cenel Augusto Perez received his PhD in Computational Linguistic from the Al. I. Cuza University of Iasi, Romania in 2014.

Radu Simionescu will receive his PhD in Computer Science from the Al. I. Cuza University of Iasi, Romania, in October 2016.

Article received on 11/01/2016; accepted 21/03/2016.
Corresponding author is Cătălina Mărănduc.