Towards the Annotation of Communicative Structure in Corpora

Alicia Burga(1), Simon Mille (1) and Leo Wanner (1,2)

(1) Department of Information and Communication Technologies
Pompeu Fabra University, C/ Roc Boronat, 138, 08018 Barcelona
(2) Catalan Institution for Research and Advanced Studies (ICREA)
<firstname>.<familyname>@upf.edu

Abstract

Communicative structure is central to the linguistic representation at nearly all levels of the Meaning-Text Models (MTMs). Its correlation with lexical and syntactic features makes it also essential for such natural language processing applications as text generation, which is about to undergo a significant shift from the symbolic, rule-based paradigm to the statistical paradigm. In the statistical paradigm, the availability of sufficiently large corpora annotated with linguistic information, and thus also with the communicative structure (CommStr), is critical. However, to the best of our knowledge, so far no corpora annotated with CommStr in the sense of the Meaning-Text Theory are available. We describe two experiments that explore how such corpora can be obtained. In the first experiment, a fragment of a Spanish Treebank is annotated manually. In the second experiment, we exploit the correlation of CommStr with syntactic features to annotate the English PropBank.

Keywords

communicative structure, MTT, treebank, annotation, corpus, Spanish, English

1 Introduction

Communicative Structure (CommStr) is central to the linguistic representation at nearly all levels of the Meaning-Text Models (MTM). As argued by Mel'čuk (2001), its various dimensions are signaled by lexical, syntactic, topological and prosodic means. For natural language processing (NLP) applications such as text generation (TG), especially the first two means are of relevance since they suggest that CommStr must be the driving instrument during lexicalization, i.e., mapping of semantemes to lexemes, and syntacticization, i.e., mapping of the Sem(antic) Str(ucture) onto the D(eep)-Synt(actic) Str(ucture) and of the DSyntStr onto the S(urface)SyntStr. Given that it is consensus among researchers that TG starts from abstract semantic or conceptual structures, one would expect CommStr to be of broad use in the field. However, this is not the case. The use of CommStr in TG is still rather

seldom. Only a restricted number of rule-based generators use it; see, e.g., (Iordanskaja et al., 1988; Wanner et al., 2003). Some works study the role of the CommStr for speech synthesis (Iomdin & Lobanov 2009; White et al., 2010; Iomdin et al. 2011), but, again, there are not many of them. This is certainly the reason why the CommStr has so far also been largely neglected in the recent statistical boom in NLP. As a consequence, hardly any corpus has been annotated with CommStr. We know just of the Prague Dependency Treebank, in which *Topic-Focus Articulation* (the equivalent of CommStr in the Prague School of linguistics) has been included (Hajič et al., 2006; Mikulová et al., 2006). This state of affairs is very unsatisfactory since our experience in statistical TG from semantic structures is that a corpus annotated with CommStr is indispensable (Bohnet et al., 2011). In what follows, we describe our ongoing work on the annotation of dependency treebanks of English and Spanish with CommStr. We explore how a large-scale annotation of the CommStr can be obtained: manually or drawing on treebanks that originally lack any communicative annotation. We believe that at least a part of the CommStr can be annotated automatically, based on the corresponding syntactic structure, but that the automatically obtained CommStr corpus should be completed by manual annotation in order to obtain fine-grained CommStrs suitable for machine-learning algorithms used in statistical NLP. The next section introduces, for convenience of the reader, the basics of the CommStr. Section 3 discusses our manual annotation exercises of the Spanish Treebank. Section 4 describes an experiment on the automatic derivation of some dimensions of CommStr from the syntactic and semantic annotation of the widely used English Treebank PropBank (Palmer et al., 2005), and Section 5 outlines our plans for future work in this area.

2 Basics of the Communicative Structure

In our interpretation of CommStr, we follow Mel'čuk (2001), who distinguishes eight communicative dimensions. The advantage of Mel'čuk's proposal is that (i) it is considerably more fine-grained than the other models of what is usually called *information structure*, and (ii) all of its dimensions are put in correlation with lexical, syntactic and prosodic means, while for information structure, only a correlation with word order and intonation has been discussed (Mikulová et al., 2006; White et al., 2010).

In what follows, we introduce the communicative dimensions that are of immediate relevance to TG and discuss how they are signaled by lexical and syntactic means particularly in the case of English and Spanish. We focus on the following five dimensions: 1. thematicity, 2. givenness, 3. focalization, 4. perspective, and 5. emphasis.[1] The dimension of **Thematicity** is given by the opposition between *rheme*, *theme* and *specifier*. *Rheme* is the content (or message) of the statement in question; *theme* marks what this message is about, and *specifier* sets the context of the message. In an English sentence, theme is, as a rule, expressed as the grammatical subject, while rheme is formed by the verbal governor with its object dependents and its local circumstantials. The sentential adverbials such as vocatives or sentential parentheticals form the specifier. **Givenness** captures the opposition between *given* and *new*. *Given* is the part of the statement that is known to the addressee, and *new* – the one that is

[1] We leave aside the dimensions of presuposedness, unitariness and locutionality because their role for generation still needs more reflection.

unknown. Gundel (1989) introduces four degrees of givenness, which correlate with different degrees of definiteness and pronominalization: *the*, *that*, *this*, and *it*. The new marker correlates with indefiniteness – *a*. **Focalization** marks parts of a statement that are in the focus of attention of the Speaker. The main means of focalization in syntax are dislocation, fronting, clefting, and conversion. **Perspective** (foregrounded vs. backgrounded vs. neutral) marks parts of the statement that are psychologically of primary / secondary relevance to the Speaker or that are not marked in terms of relevance. The main syntactic means to express *foregrounded* parts of a statement is raising; to express *backgrounded* parts, parenthetical constructions can be used. **Emphasis** deals with the emotive stress of parts of a statement. The main means to express emphasis are intonation and gestures; syntactic and lexical means include repetition (common, e.g., in Italian and Spanish) and special markers (such as the verbal *do* marker in English: *I do know what I am talking about*).

3 A first exercise in annotation of CommStr

The nature of the annotation of corpora with CommStr is different from the annotation with morphological and syntactic dependency tags since (i) CommStr tags need to be assigned to subgraphs or subtrees respectively rather than to single nodes or arcs, and (ii) the communicative tags are superimposed on the basic structure at a given level, i.e., SemStr, DSyntStr, SSyntStr, etc. In what follows, we focus on SemStr.

The first requirement for the annotation of the CommStr is to have access to the syntactic structure of the sentences to annotate. As mentioned above, in English and Spanish (as in most of the Indo-European languages), the syntactic structure directly reflects particular communicative features. In the following, we detail how to annotate the five communicative dimensions presented in Section 2. The examples cited have been gathered during the manual annotation of the CommStr on our multi-level annotated Spanish corpus. At this point, we annotated >400 sentences out of the 3.500 of the total corpus.[2] Consider Figure 1 for illustration.

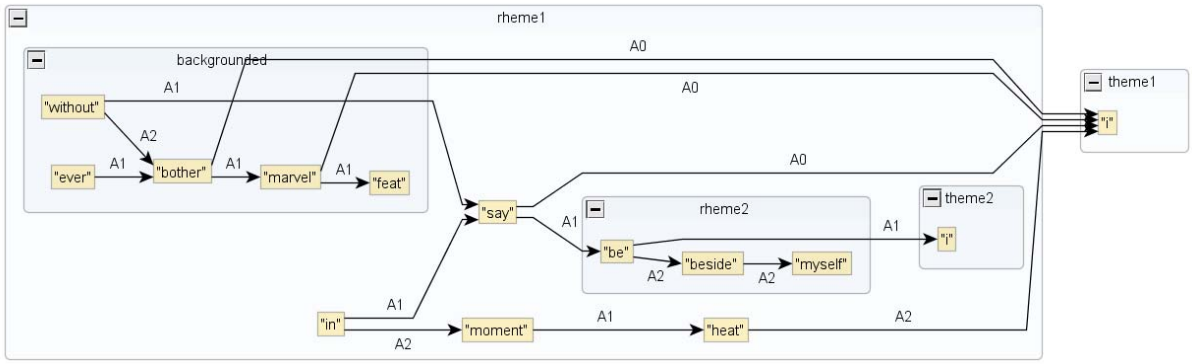

Figure 1: Sample CommStr annotation of *I've said in moments of heat, without ever bothering to marvel the feat, that I am beside myself.*

[2] For a preliminary presentation of the multilevel corpus, see (Mille et al., 2009), who use the same corpus as AnCora (Martí et al., 2008).

Thematicity: In the case of simple clauses, the annotation of thematicity is rather straightforward – except for the distinction between local and sentential adverbials, which can be problematic, especially if we talk about automatic annotation. However, thematicity is also recursive, which means that a subordinate clause (relative or completive) within a theme or rheme, and a specifier can contain their own theme and rheme. In order to differentiate the main communicative core from the embedded one(s), we use indices: theme$_{1/2/3}$…, rheme$_{1/2/3/…}$. Given that more than one embedded communicative structure is possible, the indices actually reflect the depth of each clause in the sentence.

The theme/rheme dimension can combine with any other communicative dimension, although no theme and rheme can be backgrounded as a whole. The governor of each theme and each rheme span is marked as the "main" communicative node (this information is particularly important when it comes to syntactically organize each communicative span during sentence generation).

Sentences containing indirect discourse deserve special attention due to their communicatively ambiguous interpretation: the subject is not necessarily theme nor is the main verb part of the rheme; rather, they can compose a specifier. This is the case when it is possible to replace the indirect discourse by *according to X*. Consider the sentence in (1) and its annotation. It could be interpreted – and consequently annotated – differently, if it were the case that the utterance is about the report and not about the stocks. Therefore, it is important to take into account the previous sentences (and even those that follow the sentence in question) when annotating manually, in order to evaluate whether or not the indirect discourse components are part of the thematicity core. Obviously, this kind of distinction would be extremely difficult to automate.

(1) [*El informe dijo que*]$_{Spec}$ [*las reservas de oro* […]]$_{T}$ [*eran de 18.300 millones de dólares*]$_{R}$
'The report said that the stocks of gold were 18.300 millions of dollars'.

According to Mel'čuk (2001), *wh*-words are necessarily rhematic. Following this assumption, *wh*-words are always annotated as part of the rheme. If the *wh*-word corresponds to the grammatical subject, the sentence is annotated as purely rhematic; cf. (2):

(2) ¿[*Quién* […] *puede haber diseñado una bacteria como la legionella* […]]?$_{R}$
'Who could have designed a bacterium as the legionella?'

The different components of a coordination are treated as being part of the same communicative component; see (3) – except for those cases where each component contains its own subject, as in (4).

(3) […] [*usted*]$_{T}$ [*es una persona poseída por el divino don de la caridad y quiere ayudar a sus semejantes* […]]$_{R}$
'You're a person obsessed with the divine gift of charity and want to help your fellow men.'
(4) [*El libro*]$_{T1}$ [*es divertido* […]]$_{R1}$, *y* [*su estilo*]$_{T2}$ [*un auténtico regalo*]$_{R2}$
'The book is fun and its style an authentic gift.'

In the corpus, we have found sentences with up to five themes and rhemes, although they correspond just to three levels of recursiveness; see (5) for illustration.

(5) [*Pero*]$_{Spec1}$ [*el juez*, [[*que*]$_{T2}$ […] [*dictaminó que* [*Microsoft*]$_{T5}$ [*incurrió en prácticas de monopolio*]$_{R5}$]$_{R2}$]$_{T1}$, [*opinó que* [*las medidas* […]]$_{T3}$ [*no son bastante severas*]$_{R3}$ *y se manifestó a favor de otro planteamiento* [*que*]$_{T4}$ [*dividiría a la empresa* […]]$_{R4}$]$_{R1}$
'But the judge, who ruled that Microsoft fell into monopoly practicing, expressed the view that the measures are not severe enough and declared himself in favour of another approach that would divide the firm'.

Givenness: This dimension can be marked very easily and directly: the nominal phrases that are introduced by definite articles (such as Sp. *el* and Engl. *the*) correspond to the first degree of givenness, demonstratives (such as Sp. *ese* and Engl. *that*) to the second degree, deictics (such as Sp. *este* and Engl. *this*) or possessive adjectives to the third degree. Nominal phrases headed by a pronoun are marked with the fourth degree of givenness. The rest of nominal phrases are signalled as new.

Obviously, there are no restrictions with respect to givenness when it comes to combine elements which express this dimension. In (6), the four different degrees of givenness are illustrated:

(6) a. […] *considera* [*estos cambios*]$_{G3}$ *lógicos, pues "resulta razonable que* [*las Bolsas europeas*]$_{G1}$ *unifiquen* [*sus normas de contratación*]$_{G3}$", *ya que* [*éstas*]$_{G4}$ *supondrán* ["*una ventaja* […]"]$_{N}$
'[He] considers logical these changes, since it seems reasonable that the European stock markets unify their transaction laws, given that they will mean "an advantage".'
b. […] [*la causa de* [*ese silbido*]$_{G2}$ o [*ese zumbido*]$_{G2}$]$_{G1}$ *es* [*la irritación d*[*el nervio acústico*]$_{G1}$]$_{G1}$
'The cause of that whistle or that buzzing is the irritation of the acoustic nerve.'
c. *¿No es acaso toda religión* [*la hipóstasis d*[*el conflicto entre* [*la inercia de* [*este mundo material*]$_{G3}$]$_{G1}$ *y* [*las supremas incitaciones de otro mundo*]$_{G1}$]$_{G1}$]$_{G1}$?
'Is it not every religion the hypostasis of the conflict between the inertia of this material world and the supreme incitements of another world?'

Given elements (of different degrees) can be included into other given elements (of the same or different degrees). It is also possible to find given elements within new elements (7), and vice versa (8).

(7) […] *Microsoft considera que* [*la separación de* [*la firma*]$_{G1}$]$_{G1}$ *es* [*un castigo demasiado duro para* [*las infracciones de las cuales se le acusa*]$_{G1}$]$_{N}$
'Microsoft considers the firm's division as a punishment too strong for the infractions for which it is accused.'
(8) […] *convencer a* [*la opinión pública*]$_{G1}$ *de* [*los riesgos de* [*un consumo*]$_{N}$ *y de* [*un crecimiento desbocado*]$_{N}$]$_{G1}$ […]
[…] 'to convince the public opinion of the risks of a consumption and a growth without control'.

Focalization: As mentioned in Section 2, the syntactic constructions that realize focalized content are quite obvious: fronted, cleft and promoted elements are marked as focalized.[3] In particular, adverbs and circumstantials that appear before the subject and object elements that appear before the governing verb are considered to be focalized. The rest of the sentence is marked as non-focalized (or neutral). During our annotation exercise, we have not found as

[3] The prominence of a focalized element with respect to the other elements is also reflected by intonation in that the nuclear accent is put over the focalized word (Hualde 2002). However, we do not delve into this issue here.

yet prototypical cases of dislocation or clefting, but we have found frequently cases of focalization through more subtle movements.

Focalization is directly linked to contrast. Thus, contrastive elements are focalized and marked as such; see (9) for illustration:

(9) […] *ha propuesto dar* [*a los fabricantes de ordenadores*]$_{\text{Foc}}$ *mayor flexibilidad* […] *y* [*a los consumidores*]$_{\text{Foc}}$ *más opciones* […].
'[He] has proposed to give to the computer manufacturers more flexibility and to the consumers more options.'

In order to differentiate focalized elements from foregrounded elements (which can also be marked through movement; see below), we have marked as focus those elements that contain by themselves a contrastive load:

(10) *Quiero* [*con esto*]$_{\text{Foc}}$ *decir que Medardo Fraile ha escrito un relato extraño y divertido*
lit. '[I] want with this to say that Medardo Fraile has written a strange and funny story.'

It is important to note that focalization is not recursive and focalized elements cannot appear within a backgrounded element. This information can be used when implementing a CommStr verification checker.

Perspective: Right-dislocated and parenthetical elements are marked as backgrounded. When circumstantial elements that normally appear in the periphery are located close to the verb and are not surrounded by commas, they are considered as foregrounded; other elements are marked as neutral.

Parenthetical elements are many times surrounded by commas. However, depending on the specific communicative situation, elements within commas can be interpreted as backgrounded or foregrounded. This is why we found difficulties when annotating manually this communicative dimension, and we had to turn to the sentences in the context to take a decision. When the context is of no help, we annotate by default those elements as backgrounded:

(11) *Austria conquistó 16 medallas en Salt Lake City* [*– 2 de oro , 4 de plata y 10 de bronce –*]$_{\text{Backgr}}$.
'Austria won 16 medals in Salt Lake City – 2 gold, 4 silver and 10 bronze.'

When clitics appear as markers of possession raising, they are considered foregrounded elements, equally to personal pronouns that appear before the *wh*-word in questions:

(12) […] ¿[*Tú*]$_{\text{Foregr}}$ *qué le regalarías por Reyes al duque de Feria?* […]
'What would you give to the Feria duke for Epiphany?'

Perspective is also recursive. Thus, it is possible to have foregrounded elements within backgrounded elements, as in (13), and vice versa, as in (14).

(13) *El gobernante*, [*con ganada fama* [*desde que llegó hace 16 meses al poder*]$_{\text{Foregr}}$ *de explotar al máximo su oratoria* […]]$_{\text{Backgr}}$, *enmudeció* […]
'The leader, with earned reputation since he got 16 months ago the power of exploiting the most his oratory, fell silent.'

(14) *Los últimos dos meses* […] *han estado marcados por insultos personales,* [*principalmente entre Labastida y Fox,* [*quienes se han dicho desde "mariquita", "feo"* […], [*entre otros calificativos*]$_{\text{Backgr}}$]$_{\text{Backgr}}$]$_{\text{Foregr}}$.
'The last two months have been marked by personal insults, especially between Labastida and Fox, who have said to each other from "wimp", "ugly", among other words.'

Appositions are another aspect that deserves discussion with respect to perspective. Even if appositions do not form backgrounded or foregrounded elements (Mel'čuk, personal communication), they do seem to play the role of a psychological relevance marker. We currently explore the precise nature of this marker.

Emphasis: As mentioned in Section 2, some lexical markers express an emotional load and thus emphasis. Although it is impossible to offer a comprehensive list of those markers (given that it is the context which finally defines whether or not an element is emphatic), we can make rough but useful generalizations. Thus, we annotate as emphatic some occurrences of the adverbial *una vez más* 'once again', *también* 'too', *muy* 'very', as well as superlative adjectives and adverbs (15). Which of them are in fact emphatic is decided upon the analysis of each occurrence.

(15) […] *procuraría que el regalo, además de* [*carísimo*]$_{\text{Emph}}$*, tuviera directamente que ver con el mayor vicio del obsequiado.*
'[I]'d try to make sure that the gift, as well as being expensive, it's also directly related to the greatest vice of who receives it.'

Repetition (approximate or exact, total or partial) of some elements is a syntactic means to emphasize a part of the utterance. We mark as emphasized the first element (in that we assume that it is emphasized through repetition) and as marker of emphasis the repetition itself:[4]

(16) *El libro es* [*divertido*]$_{\text{Emph}}$, [*muy divertido*]$_{\text{Emph\_Marker}}$ […].
'The book is fun, very fun.'

Emphasis is neither recursive nor obligatory. Even though theoretically emphasis can be combined with any other communicative dimension, so far we have found in the corpus emphasis combined with thematicity and givenness.

In addition to the five communicative dimensions discussed above, we annotate parts of utterances which appear within quotation marks at the surface with the "signalled" tag of the locutionality dimension.[5] Otherwise, it would not be possible to use quotation marks in statistical generation.

---

[4] We are using here the term 'repetition', but another more appropriate term could be proposed, given that sometimes the "repetition" consists on making explicit some semantic characteristics of the term to be emphasized, as in (i):
(i) […] [*titula*]$_{\textit{Emph}}$ *[en portada]*$_{\textit{Emph\_marker}}$ *"Villalonga normaliza las relaciones* […]*"*
'(It) heads in the front page "Villalonga normalizes the relations […]'

[5] According to Mel'čuk (2001), locutionality distinguishes between "communicating" and "signalling" utterances. The first ones pretend to explicitly communicate something and, in that sense, they can be headed by the phrase "I want you to know that…". The second ones just signal something that happens inside the speaker

4 Deriving CommStr from a Syntactic Dependency Annotation

The fact that part of the CommStr is signalled by lexical and syntactic means led us launch an experiment on the derivation of the CommStr from a syntactic dependency corpus annotation. The experiment has been performed on the dependency variant of the Penn Treebank (PTB) / PropBank (PB) corpus (Palmer et al., 2005), one of the most widely used corpora for English since it has been released in the standard corpus annotation format CoNLL (Hajič 2009), which contains in the same data structure syntactic and semantic annotations.

4.1 Experiment of the derivation of CommStr

The automatic derivation of the CommStr from the PTB/PB annotations has been performed using the rule-based MATE graph transducer (Bohnet & Wanner, 2010). The derivation is based on a set of rules which use semantic and superficial syntactic and topological criteria available in PB and PTB. For instance, the subject of a sentence is marked as theme and the corresponding VP as rheme; an indefinite NP is marked as new; and so on.

The available criteria are, obviously, too crude and too simplistic to capture the information structure in its entirety. The fact that the CommStr–SyntStr projection is not isomorphic makes the derivation even more difficult. Therefore, we focused on the derivation of the opposition theme vs. rheme (ignoring the feature of Specifier), the dimension of Givenness and a combined version of Perspective and Focalization we called "Foregroundedness".

4.2 Assessment of the derivation

Despite the limited range of criteria we could use for the derivation of the partial CommStr, the evaluation below shows that the obtained CommStr may well serve as a first approximative annotation.

4.2.1 Quantitative Evaluation

To assess the quality of the derivation, we performed a quantitative evaluation in which we compared the automatically obtained CommStr with a gold standard of 90 sentences.[6] Table 1 shows the results of our quantitative evaluation:'thematicity *p*/*r*' stands for precision and recall of the theme/rheme introduction; 'main *p*/*r*' for precision and recall of the marking of the main node of all themes and rhemes, and 'th-rh pairs *p*/*r*' for precision and recall of the identification of the theme/rheme alignment (each theme is marked as being the theme of a particular rheme of the sentence). 'foregr/backgr *p*/r' stands for the accuracy of the perspective annotation (foregrounded/focalized vs. backgrounded vs. neutral), 'depth *p*/*r*' for

(they do not express linguistically the communication act), and in that sense they cannot be negated or questioned.

[6] We are fully aware that 90 sentences are not sufficient to objectively assess the results of our experiment. However, even with such a small gold standard corpus it is possible to estimate whether the adopted strategy is promising or not.

precision and recall of the recursive theme/rheme (primary, secondary, terciary, etc.) annotation, and 'given *p*/*r*' for the accuracy of givenness annotation.

thematicity		main		th-rh pairs		foregr/backgr		depth		given	
p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
0.986	1.0	1.0	0.951	0.914	1.0	0.905	0.358	0.807	1.0	1.0	0.986

Table 1: Precision and recall for the automatic introduction of CommStr dimensions

The numbers show that the identification of the main node, theme/rheme and givenness works well. This is because these notions actually correlate very much with some prominent syntactic features that could be deduced from the PTB annotation. The accuracy of the annotation of the recursive theme/rheme structure is somewhat lower. This is because every node receives its thematicity feature from the main node of the span it belongs to, but each node can have more than one governor, and each governor can belong to a different communicative span. The recall of the assignment of the perspective is rather low (0.358), although the precision is high (0.905). This means that syntactic and topological clues only are by far not sufficient to determine which element is to be marked as foregrounded, which one as backgrounded, and which one is neutral with respect to communicative prominence. Other types of clues are also needed.

thematicity			main			th-rh pairs			foregr/backgr			depth			given		
tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn	tp	fp	fn
704	10	0	135	0	7	64	6	0	38	4	68	569	136	0	70	0	1

Table 1: Total numbers of the quality of the annotation of the individual communicative features ('tp' stands for "true positives", 'fp' for "false positives", and 'fn" for "false negatives")

4.2.2 Limitations of the derivation of CommStr from syntactic annotation

The automatic derivation of CommStr from a syntactic annotation can only be partial. First of all, in Indo-European languages, there are few distinctive syntactic constructions for focalization, mainly clefting and left dislocation. However, a left dislocation can be difficult to interpret, since it can also correspond to neutrality from the point of view of focalization (consider, e.g., *yesterday* in *Yesterday, I went to the beach.*). Emphasis is often spotted thanks to the presence of particular cue words in a particular position. For perspective, the presence of parentheses is a clear marker of backgroundedness, but as far as the other features are concerned, it is necessary to look at the positioning of the groups. The fact that importance is given to the ordering among the components of a sentence is also a problem by itself: it raises issues when it comes to languages with free word order, such as Russian, for instance. In addition, some reasoning is necessary in order to interpret a sentence; an algorithm will probably never be able to recognize a slightly peculiar construction or a combination of words which are intended to signal a particular communicative goal of the speaker.

5 Conclusions and future work

The rule-based derivation of CommStr from syntactic annotation such as PennTreeBank and PropBank is an option that can be considered to ensure a short term availability of a corpus annotated with CommStr. However, if a high quality, detailed annotation is targeted, machine learning (ML) based annotation seems more adequate. Given that ML-based annotation requires manually annotated corpora as training material, we need to enlarge our corpora, as well as to guarantee the quality of the annotation compiling precise guidelines for the annotators, using such metrics as inter-annotator agreement and foreseeing a posterior revision iteration. But in order to be able to compile precise annotation guidelines we still need to discuss and decide how to treat certain phenomena, such as the distinction between focalized and foregrounded elements, or the definition of emphatic markers.

Acknowledgements

Thanks to Bernd Bohnet for his help with the derivation of CommStr from the PropBank annotation and the subsequent evaluation. This work has been partially supported by the European Commission under the contract number FP7-ICT-248594, by the Spanish Ministry of Science and Innovation under the contract number FFI2008-06479-C02-02 and by the Funds FEDER of the European Commission.

Bibliography

Bonhet, B. & L. Wanner. 2010. Open Source Graph Transducer and Interpreter Development Environment. In Proceedings of the LREC.

Bonhet, B., L. Wanner, & S. Mille. 2011. Statistical Language Generation from Semantic Structures. In Proceedings of the DepLing.

Gundel, J., N. Hedberg, & R. Zacharski. 1989. Giveness, Implicature and Demonstrative Expressions in English Discourse. In CLS-25, Part II (Parasession on Language in Context), pages 89–103. Chicago Linguistics Society.

Hajič, J. et al. 2006. *Prague Dependency Treebank 2.0* . In Linguistic Data Consortium, Philadelphia.

Hajič, J. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Proceedings of the CoNLL.

Hualde, J.I. 2002. Intonation in Spanish and the other Ibero-Romance languages: Overview and status quaestionis. In *Romance Phonology and Variation. Selected Papers from the 30th Linguistic Symposium on Romance Languages, Gainesville, Florida, February 2000*, ed. by Caroline Wiltshire and Joaquim Camps, 101-115. Amsterdam: Benjamins.

Iomdin, L.L. & B.M. Lobanov. 2009. Syntactic Correlates of Prosodically Marked Elements of the Sentence and Their Role in the Tasks of Text-to-Text Speech Synthesis. In Proceedings of the Dialog '09 Conference.

Iomdin, L.L., B.M. Lobanov & Ju.S. Gecevich. 2011. The Talking ETAP. Using the ETAP Parser in Russian Speech Synthesis. In Proceedings of the Dialog '11 Conference.

Iordanskaja, L. N., R. Kittredge & A. Polguère. 1988. Implementing a Meaning-Text Model for Language Generation. In *Proceedings of COLING 1988*.

Martí, M.A., M. Taulé, L. Márquez, & M. Bertran. 2008. Ancora: A Multilingual and Multilevel Annotated Corpus, Pending to be published (http://clic.ub.edu/corpus/ancora-publicacions)

Mikulová, M. et al. 2006. *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Reference Book.* http://ufal.mff.cuni.cz/pdt2.0update/doc/tr-ref-cz-en.pdf

Mel'čuk, I.A. 2001. *Communicative Organization in Natural Language : The Semantic-Communicative Structure of Sentences.* John Benjamins Publishing, Philadelphia.

Mille, S., Burga, A., Vidal, V. & Wanner, L. 2009. "Towards a Rich Dependency Annotation of Spanish Corpora". In *Proceedings of SEPLN'*09, San Sebastian.

Palmer, Martha, D. Gildea, & P. Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles, in *Computational Linguistics Journal*, 31:1.

Wanner, L, B. Bohnet, & M. Giereth. 2003 Deriving the Communicative Structure in Applied NLG. in *Proceedings of the 9th European Natural Language Generation Workshop* at the Annual Meeting of the Association for Computational Linguistics, Budapest, 111-118.

White, M, R.A.J. Clark, & J. Moore. 2010 Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, 36(2):159–201.