Towards a Surface Realization-Oriented Corpus Annotation

Leo Wanner
ICREA and
Universitat Pompeu Fabra
Roc Boronat 138
Barcelona, 08018, Spain
leo.wanner@upf.edu

Simon Mille
Universitat Pompeu Fabra
Roc Boronat 138
Barcelona, 08018, Spain
simon.mille@upf.edu

Bernd Bohnet
Universität Stuttgart
IMS, Pfaffenwaldring 5b
Stuttgart, 70569, Germany
bohnet@ims.uni-stuttgart.de

Abstract

Until recently, deep stochastic surface realization has been hindered by the lack of semantically annotated corpora. This is about to change. Such corpora are increasingly available, e.g., in the context of CoNLL shared tasks. However, recent experiments with CoNLL 2009 corpora show that these popular resources, which serve well for other applications, may not do so for generation. The attempts to adapt them for generation resulted so far in a better performance of the realizers, but not yet in a genuinely semantic generation-oriented annotation schema. Our goal is to initiate a debate on how a generation suitable annotation schema should be defined. We define some general principles of a semantic generation-oriented annotation and propose an annotation schema that is based on these principles. Experiments shows that making the semantic corpora comply with the suggested principles does not need to have a negative impact on the quality of the stochastic generators trained on them.

1 Introduction

With the increasing interest in data-driven surface realization, the question on the adequate annotation of corpora for generation also becomes increasingly important. While in the early days of stochastic generation, annotations produced for other applications were used (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998; Bangalore and Rambow, 2000; Oh and Rudnicky, 2000; Langkilde-Geary, 2002), the poor results obtained, e.g., by (Bohnet et al., 2010) with the original CoNLL 2009 corpora, show that annotations that serve well for other applications, may not do so for generation and thus need at least to be adjusted.[1] This has also been acknowledged in the run-up to the surface realization challenge 2011 (Belz et al., 2011), where a considerable amount of work has been invested into the conversion of the annotations of the CoNLL 2008 corpora (Surdeanu et al., 2008), i.e., PropBank (Palmer et al., 2005), which served as the reference treebank, into a more "generation friendly" annotation. However, all of the available annotations are to a certain extent still syntactic. Even PropBank and its generation-oriented variants contain a significant number of syntactic features (Bohnet et al., 2011b).

Some previous approaches to data-driven generation avoid the problem related to the lack of semantic resources in that they use hybrid models that imply a symbolic submodule which derives the syntactic representation that is then used by the stochastic submodule (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998). (Walker et al., 2002), (Stent et al., 2004), (Wong and Mooney, 2007), and (Mairesse et al., 2010) start from deeper structures: Walker et al. and Stent et al. from *deep-syntactic structures* (Mel'čuk, 1988), and Wong and Mooney and Mairesse et al. from higher order predicate logic structures. However, to the best of our knowledge,

[1] Trained on the original ConLL 2009 corpora, (Bohnet et al., 2010)'s SVM-based generator reached a BLEU score of 0.12 for Chinese, 0.18 for English, 0.11 for German and 0.14 for Spanish. Joining the unconnected parts of the sentence annotations to connected trees (as required by a stochastic realizer) improved the performance to a BLEU score of 0.69 for Chinese, 0.66 for English, 0.61 for German and 0.68 for Spanish.

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 22–30,
Utica, May 2012.

none of them uses corpora annotated with the structures from which they start.

To deep stochastic generation, the use of hybrid models is not an option and training a realizer on syntactically-biased annotations is highly problematic in the case of data-to-text NLG, which starts from numeric time series or conceptual or semantic structures: the syntactic features will be simply not available in the input structures at the moment of application.[2]. Therefore, it is crucial to define a theoretically sound semantic annotation that is still good in practical terms.

Our goal is thus to discuss some general principles of a semantic generation-oriented annotation schema and offer a first evaluation of its possible impact on stochastic generation. Section2 details what kind of information is available respectively not available during data-to-text generation. Section 3 states some general principles that constrain an adequate semantic representation, while Section 4 formally defines their well-formedness. Section 5 reports then on the experiments made with the proposed annotation, and Section6 offers some conclusions.

2 What can we and what we cannot count on?

In data-to-text or ontology-to-text generation, with the standard *content selection–discourse structuring—surface generation* pipeline in place, and no hard-wired linguistic realization of the individual chunks of the data or ontology structure, the input to the surface realization module can only be an abstract structure that does not contain any syntactic (and even lexical) information. *Conceptual graphs* in the sense of Sowa (Sowa, 2000) are structures of this kind;[3] see Figure 1 for illustration ('Cmpl' = 'Completion', 'Rcpt' = 'Recipient', 'Strt' = 'Start', 'Attr' = Attribute, 'Chrc' = 'Characteristic', and 'Amt' = 'Amount'). *Content selection* accounts for the determination of the content units that are to be communicated and *Discourse Structuring* for the delimitation of *Elementary Discourse Units* (EDUs) and their organization and for the discursive relations between them (e.g., *Bcas* (*Because*) in the Figure).

In particular, such a structure cannot contain:

- non-meaningful nodes: governed prepositions (BECAUSE *of*, CONCENTRATION *of*), auxiliaries (passive *be*), determiners (*a*, *the*);
- syntactic connectors (*between* A *and* B), relative pronouns, etc.
- syntactic structure information: A modifies B, A is the subject of B, etc.

In other words, a deep stochastic generator has to be able to produce all syntactic phenomena from generic structures that guarantee a certain flexibility when it comes to their surface form (i.e., without encoding directly this type of syntactic information). For instance, *a concentration of NO2* can be realized as *a NO2 concentration*, *between 23h00 and 00h00* as *from 23h00 until 00h00*, etc. This implies that deep annotations as, for instance, have been derived so far from PennTreeBank/PropBank, in which either all syntactic nodes of the annotation are kept (as in (Bohnet et al., 2010)) or only certain syntactic nodes are removed (as THAT complementizers and TO infinitives in the shared task 2011 on surface realization (Belz et al., 2011)) still fall short of a genuine semantic annotation. Both retain a lot of syntactic information which is not accessible in genuine data-to-text generation: nodes (relative pronouns, governed prepositions and conjunctions, determiners, auxiliaries, etc.) and edges (relative clause edges, control edges, modifier vs. argumental edges, etc.).

This lets us raise the question how the annotation policies should look like to serve generation well and to what extent existing resources such as PropBank comply with them already. We believe that the answer is critical for the future research agenda in generation and will certainly play an outstanding role in the shared tasks to come.

In the next section, we assess the minimal principles which the annotation suitable for (at least) data-to-text generation must follow in order to lead to a *core semantic structure*. This core structure still ignores such important information as co-reference,

[2]Even though in this article we are particularly interested in data-to-text generation, we are convinced that clean semantic and syntactic annotations also facilitate text-to-text generation.

[3]But note that this can be any other content structure.

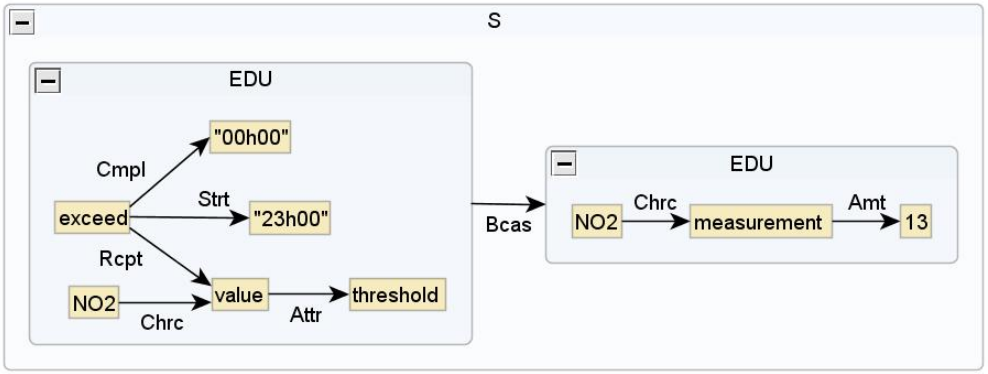

Figure 1: Sample conceptual structure as could be produced by text planning (*Because of a concentration of NO2 of 13$\mu g/m^3$, the NO2 threshold value was exceeded between 23h00 and 00h00*)

scope, presupposition, etc.: this information is obviously necessary, but it is not absolutely vital for a sufficient restriction of the possible choices faced during surface generation. Further efforts will be required to address its annotation in appropriate depth.

3 The principles of generation-suitable semantic annotation

Before talking about generation-suitable annotation, we must make some general assumptions concerning NLG as such. These assumptions are necessary (but might not always be sufficient) to cover deep generation in all its subtleties: (i) data-to-text generation starts from an abstract conceptual or semantic representation of the content that is to be rendered into a well-formed sentence; (ii) data-to-text generation is a series of equivalence mappings from more abstract to more concrete structures, with a chain of inflected words as the final structure; (iii) the equivalence between the source structure S_s and the target structure S_t is explicit and self-contained, i.e., for the mapping from S_s to S_t, only features contained in S_s and S_t are used. The first assumption is in the very nature of the generation task in general; the second and the third are owed to requirements of statistical generation (although a number of rule-based generators show these characteristics as well).

The three basic assumptions give rise to the following four principles.

1. Semanticity: The semantic annotation must capture the meaning and only the meaning of a given sentence. Functional nodes (auxiliaries, determiners, governed conjunctions and prepositions), node duplicates and syntactically-motivated arcs should not appear in the semantic structure: they reflect grammatical and lexical features, and thus already anticipate how the meaning will be worded. For example, *meet*-AGENT→*the* (*directors*), *meet*-LOCATION→*in* (*Spain*), *meet*-TIME→*in* (*2002*) cited in (Buch-Kromann et al., 2011) as semantic annotation of the phrase *meeting between the directors in Spain in 2002* in the Copenhagen Dependency Treebank does not meet this criterion: *the*, and both *in*s are functional nodes. Node duplicates such as the relative pronoun *that* in the PropBank annotation (*But Panama illustrates that their their substitute is a system*) *that*←R-A0-*produces* (*an absurd gridlock*) equally reflect syntactic features, as do syntactically-motivated arc labels of the kind 'R(elative)-A0'.

The PropBank annotation of the sentence cited above also intermingles predicate-argument relations ('Ai') with syntactico-functional relations ('AM-MNR'): *gridlock*−AM-MNR→*absurd*. The predicate-argument analysis of modifiers suggests namely that they are predicative semantemes that take as argument the node that governs them in the syntactic structure; in the above structure: $absurd-$A1$\rightarrow gridlock$. This applies also to locatives, temporals and other "circumstantials", which are most conveniently represented as two-place semantemes: $house\leftarrow$A1$-location-$A2$\rightarrow Barcelona$, $party\leftarrow$A1$-time-$A2$\rightarrow yesterday$, and so on. Although not realized at the surface, $location$, $time$, etc. are crucial.

2. Informativity: A propositional semantic annotation must be enriched by *information structure* features that predetermine the overall syntactic structure (paratactic, hypotactic, parenthetical, ...), the internal syntactic structure (subject/object, clefted or not, any element fronted or not, etc.), determiner distribution, etc. in the sentence. Otherwise, it will be always underspecified with respect to its syntactic equivalence in that, as a rule, a single semantic structure will correspond to a number of syntactic structures. This is not to say that with the information structure in place we will always achieve a 1:1 correspondence between the semantic and syntactic annotations; further criteria may be needed—including prosody, style, presupposedness, etc. However, information structure is crucial.

The most relevant information structure features are those of Thematicity, Foregroundedness and Givenness.[4]

Thematicity specifies what the utterance states (marked as *rheme*) and about what it states it (marked as *theme*).[5] Theme/rheme determines, in the majority of cases, the subject-object structure and the topology of the sentence. For instance,[6] $[John]_{theme}$←A1−$[see$−A2→$Maria]_{rheme}$ may be said to correspond to $John$←subject−see−dir.obj→$Maria$ and $[John$←A1−$see]_{rheme}$−A2 →$[Maria]_{theme}$ to $John$ ←obj−see_{pass}−subject →$Maria$. For the generation of relative sentence structures such as *John bought a car which was old and ugly*, we need to accommodate for a recursive definition of thematicity: $[John]_{theme}$←A1−$[buy$−A2→$[c1 : car]_{theme}$←A1−$[old]_{rheme}$; $c1$←A1 −$[ugly]_{rheme}]_{rheme}$.[7] With no recursive (or *secondary* in Mel'čuk's terms) thematicity, we would get *John bought an old and ugly car*.[8]

It is quite easy to find some counter-examples to the default theme/rheme–syntactic feature correlation, in particular in the case of questions and answers. For instance, the neutral answer to the question *What will John bake tomorrow?*, *John will bake a cake*, would be split as follows: $[John$←A1−$bake]_{theme}$ −A2→$[cake]_{rheme}$. In this case, the main verb at the surface, *bake*, is included in the theme and not in the rheme. Consider also the sentence *In a cross-border transaction, the buyer is in a different region of the globe from the target*, where the main theme is *in a cross-border transaction*, i.e., not the subject of the sentence (with the subject *the buyer* being the embedded theme of the main rheme). In these cases, the correlation is more complex, but it undoubtedly exists and needs to be distilled during the training phase.

Foregroundedness captures the "prominence" of the individual elements of the utterance for the speaker or hearer. An element is 'foregrounded' if it is prominent and 'backgrounded' if it is of lesser prominence; elements that are neither foregrounded nor backgrounded are 'neutral'. A number of correlations can be identified: (i) a 'foregrounded' A1 argument of a verb will trigger a clefting construction; e.g., $[John]_{foregr;theme}$←A1−$[see$−A2→$Maria]_{rheme}$ will lead to *It was John who saw Maria*; similarly, $[John$←A1−$bake]_{foregr;theme}$ −A2→$[cake]_{rheme}$ will lead to *What John will bake is a cake*; (ii) a 'foregrounded' A2 argument of a verb will correspond to a clefting construction or a dislocation: *It was Maria, whom John saw*; (iii) a 'foregrounded' A1 or A2 argument of a noun will result in an *argument promotion*, as, e.g., *John's arrival* (instead of *arrival of John*); (iv) a 'foregrounded' circumstantial will be fronted: *Under this tree he used to rest*; (v) marking a part of the semantic structure as 'backgrounded' will lead to a parenthetical construction: *John (well known among the students and professors alike) was invited as guest speaker*. If no elements

[4] We use mainly the terminology and definitions (although in some places significantly simplified) of (Mel'čuk, 2001), who, to the best of our knowledge, establishes the most detailed correlation between information structure and syntactic features.

[5] Similar notions are *topic/focus* (Sgall et al., 1986) and *topic/comment* (Gundel, 1988).

[6] As in PropBank, we use 'Ai' as argument labels of predicative lexemes, but for us, 'A1' stands for the first argument, 'A2' for the second argument, etc. That is, in contrast to PropBank, we do not support the use of 'A0' to refer to a lexeme's external argument since the distinction between external and internal arguments is syntactic.

[7] $c1$ is a "handle" in the sense of *Minimal Recursion Semantics* (Copestake et al., 1997).

[8] We believe that operator scopes (e.g., negations and quantifiers) can, to a large extent, be encoded within the thematic structure; see (Cook and Payne, 2006) for work in the LFG-framework on German, which provides some evidence for this. However, it must be stated that very little work has been done on the subject until now.

are marked as foregrounded/backgrounded, the default syntactic structure and the default word order are realized.

Givenness captures to what extent an information element is present to the hearer. The elementary givenness markers 'given' and 'new' correlate in syntax with determiner distribution. Thus, the 'new' marker of an object node will often correspond to an indefinite or zero determiner of the corresponding noun: *A masked man was seen to enter the bank* (*man* is newly introduced into the discourse). The 'given' marker will often correlate with a definite determiner: *The masked man* (whom a passerby noticed before) *was seen to enter the bank*. To distinguish between demonstratives and definite determiners, a gradation of givenness markers as suggested by Gundel et al. (Gundel et al., 1989) is necessary: 'given$_{1/2/3}$'.

As already for Thematicity, numerous examples can be found where the giveness-syntactic feature correlation deviates from the default correlation. For instance, in *I have heard a cat, the cat of my neighbour*, there would be only one single (given) node *cat* in the semantic structure, which does not prevent the first appearance of *cat* in the sentence to be indefinite. In *A warrant permits a holder that he acquire one share of common stock for $17.50 a share*, *warrant* is given, even if it is marked by an indefinite determiner. Again, this only shows the complexity of the annotation of the information structure, but it does not call into question the relevance of the information structure to NLG.

As one of the few treebanks, the Prague Dependency Treebank (PDT) (Hajič et al., 2006) accounts for aspects of the information structure in that it annotates *Topic-Focus Articulation* in terms of various degrees of *contextual boundness*, which are correlated with word order and intonation (Mikulová et al., 2006, p.152).

3. Connectivity: The semantic annotation must ensure that the annotation of an utterance forms a connected structure: without a connected structure, generation algorithms that imply a traversal of the input structure will fail to generate a grammatical sentence. For instance, the PropBank annotation of the sentence *But Panama illustrates that their substitute is a system that produces an absurd gridlock* (here shown partially) does not comply with this principle since it consists of four unconnected meaning-bearing substructures (the single node 'but' and the subtrees governed by 'illustrate', 'produce' and 'substitute'): *but* | *Panama*←A0−*illustrate*−A1→*that* | *system*←A0−*produce*−A1→*gridlock*−AM-MNR→*absurd* | *substitute*−A0→*their*.

4 Outline of a Generation-Oriented Annotation

The definitions below specify the syntactic well-formedness of the semantic annotation. They do not intend to and cannot substitute a detailed annotation manual, which is indispensable to achieve a semantically accurate annotation.

Definition 1: [Semantic Annotation of a sentence S, SA]
SA of S in the text T in language $\mathcal{L}$ is a pair $\langle S_{sem}, S_{inf} \rangle$, where S_{sem} is the semantic structure of S (ensuring Semanticity and Connectivity), and S_{inf} is the information structure of S (ensuring Informativity).

Let us define each of the two structures of the semantic annotation in turn.

Definition 2: [Semantic Structure of a sentence S, S_{sem}]
S_{sem} of S is a labeled acyclic directed connected graph (V, E, γ, λ) defined over the vertex label alphabet $L := L_S \cup M_C \cup M_T \cup M_t \cup M_a$ (such that $L_S \cap (M_C \cup M_T \cup M_t \cup M_a) = \emptyset$) and the edge label alphabet $R_{sem} \subseteq \{$A1, A2, A3, A4, A5, A6$\}$, with
– V as the set of vertices;
– E as the set of directed edges;
– γ as the function that assigns each $v \in V$ an element $l \in L$;
– λ as the function that assigns each $e \in E$ an element $a \in R_{sem}$;
– L_S as the meaning bearing lexical units (LUs) of S;
– $M_C \subseteq \{$LOC, TMP, EXT, MNR, CAU, DIR, SPEC, ELAB, ADDR$\}$ as the "circumstantial meta semantemes" (with the labels standing for 'locative', 'temporal', 'temporal/spatial extension', 'manner', 'cause', 'direction', 'specification', 'elaboration', and 'addressee');
– $M_T \subseteq \{$TIME, TCST$\}$ as the "temporal meta semantemes" (with the labels standing for 'time' and

'time constituency');
– $M_t \subseteq \{\texttt{past}*, \texttt{present}*, \texttt{future}*\}$ as the "time value semantemes";
– $M_a \subseteq \{\texttt{imperfective}*, \texttt{durative}*, \texttt{semelfactive}*, \texttt{iterative}*, \texttt{telic}*, \texttt{atelic}*, \texttt{nil}*\}$ as the "aspectual value semantemes"[9]

such that the following conditions hold:

(a) the edges in S_{sem} are in accordance with the valency structure of the lexical units (LUs) in S: If l_p−Ai→$l_r \in S_{sem}$ ($l_p, l_r \in L_S$, $i \in \{1, 2, 3, \ldots\}$), then the semantic valency of l_p possesses at least i slots and l_r fulfils the semantic restrictions of the i-th slot

(b) the edges in S_{sem} are exhaustive: If $\gamma(n_r) = l_r \in L$ instantiates in S the i-th semantic argument of $\gamma(n_p) = l_p$, then l_p−Ai→$l_r \in S_{sem}$

(c) S_{sem} does not contain any duplicated argument edges: If $\gamma(n_p)$−Ai→$\gamma(n_r)$, $\gamma(n_p)$ −Aj→ $\gamma(n_q) \in S_{sem}$ (with $n_p, n_r, n_q \in N$) then Ai ≠ Aj and $n_r \neq n_q$

(d) circumstantial LUs in S are represented in S_{sem} by two-place meta-semantemes: If $l_r \in L_{sem}$ is a locative/temporal/ manner/cause/direction/specification/elaboration/addressee LU and in the syntactic dependency structure of S, l_r modifies l_p, then l_r←A2-α-A1→$l_p \in S_{sem}$ (with $\alpha \in$ LOC, TMP, MNR, CAU, DIR, SPEC, ELAB, ADDR)

(e) verbal tense is captured by the two-place predicate TIME: If $l_p \in L_{sem}$ is a verbal LU then l_r←A2-TIME-A1→$l_p \in S_{sem}$, with $l_r \in M_t$

(f) verbal aspect is captured by the two-place predicate TCST: If $l_p \in L_{sem}$ is a verbal LU then l_r←A2-TCST-A1→$l_p \in S_{sem}$, with $l_r \in M_a$.

(a) implies that no functional node is target of an argument arc: this would contradict the semantic valency conditions of any lexeme in S. (b) ensures that no edge in S_{sem} is missing: if a given LU is an argument of another LU in the sentence, then there is an edge from the governor LU to the argument LU. (c) means that no predicate in S_{sem} possesses in S two different instances of the same argument slot. The circumstantial meta-semantemes in (d) either capture the semantic role of a circumstantial that would otherwise get lost or introduce a predicate type for a name. Most of the circumstantial meta-semantemes reflect PropBank's modifier relations 'AM-X' (but in semantic, not in syntactico-functional terms), such that their names are taken from PropBank or are inspired by PropBank. LOC takes as A1 a name of a location of its A2: *Barcelona* ←A1-LOC-A2→*live*-A1→*John*; TMP a temporal expression: *yesterday* ←A1-TMP-A2→*arrive*-A1→*John*; MNR a manner attribute: *player*←A1-MNR-A2→*solo*; CAU the cause: *accept*←A1-CAU-A2→*reason* in *This is the reason why they accepted it*; DIR a spatial direction: *run around*←A2-DIR-A1→*circles* in *I'm running around in circles*; SPEC a "context specifier": *should*←A2-SPEC-A1→thought in *You should leave now, just a thought*; ELAB an appositive attribute *company*←A1-ELAB-A2 →*bank* in *This company, a bank, closed*; and ADDR direct address: *come*←A1-ADDR-A2→*John* in *John, come here!*

Definition 3: [Information Structure of a sentence S, S_{inf}]

Let S_{sem} of S be defined as above. S_{inf} of S is an undirected labeled hypergraph (V, I) with V as the set of vertices of S and I the set of hyperedges, with $I :=$ {theme$_i$ (i = 1, 2, ...), rheme$_i$ (i = 1, 2, ...), given$_j$ (j = 1,...,3), new, foregrounded, backgrounded}. The following conditions apply:

(a) thematicity is recursive, i.e., a thematic hyperedge contains under specific conditions embedded theme/rheme hyperedges: If $\exists n_k \in$ theme$_i$ such that $\gamma(n_k) = l_p$ is a verbal lexeme and l_p -A1→$l_r \in S_{sem}$, then $\exists$ theme$_{i+1}$, rheme$_{i+1} \in$ theme$_i$

(b) theme and rheme hyperedges of the same recursion level, given and new hyperedges, and foregrounded and backgrounded hyperedges are disjoint: theme$_i \cap$ rheme$_i = \emptyset$ (i = 1, 2, ...), given$_j$ $\cap$ new $= \emptyset$ (j = 1,...,3), foregr. $\cap$ backgr. $= \emptyset$

(c) any node in S_{sem} forms part of either theme or rheme: $\forall n_p \in S_{sem} : n_p \in$ theme$_1 \cup$ rheme$_1$.

Consider in Figure 2 an example of SA with its two structures.[10] All syntactic nodes have been removed, and all the remaining nodes are connected in terms of a predicate–argument structure, with no use of any syntactically motivated edge, so as to ensure that the structure complies with the Semanticity and Connectivity principles. Figure 2 illustrates the three main aspects of Informativity: (i) thematic-

[9] The aspectual feature names are mainly from (Comrie, 1976).

[10] The meta-semanteme TCST is not shown in the figure.

ity, with the two theme/rheme oppositions; (ii) foregroundedness, with the backgrounded part of the primary rheme; and (iii) givenness, with the attribute *givenness* and the value *2* on the node *program*. The information structure constrains the superficial realization of the sentence in that the primary theme will be the subject of the sentence, and the main node of the primary rheme pointing to it will be the main verb of the same sentence. The secondary theme and rheme will be realized as an embedded sentence in which *you* will be the subject, that is, forcing the realization of a relative clause. However, it does not constrain the appearance of a relative pronoun. For instance: *we obtained technologies you do not see anywhere else* and *we obtained technologies that you do not see anywhere else* are possible realizations of this structure. Leaving the relative pronoun in the semantic structure would force one realization to occur when it does not have to (both outputs are equally correct and meaning-equivalent to the other). Similarly, marking *the Soviet space program* as backgrounded leaves some doors open when it comes to surface realization: *Cosmos, the Soviet space program* vs. *Cosmos* (*the Soviet space program*) vs. *the Soviet space program Cosmos* (if *Cosmos* is backgrounded too) are possible realizations of this substructure.

ELABORATION is an example of a meta-node needed to connect the semantic structure: *Cosmos* and *program* have a semantic relation, but neither is actually in the semantic frame of the other—which is why the introduction of an extra node cannot be avoided. In this case, we could have a node *NAME*, but *ELABORATION* is much more generic and can actually be automatically introduced without any additional information.

5 Experiments

Obviously, the removal of syntactic features from a given standard annotation, with the goal to obtain an increasingly more semantic annotation, can only be accepted if the quality of (deep) stochastic generation does not unacceptably decrease. To assess this aspect, we converted automatically the PropBank annotation of the WSJ journal as used in the CoNLL shared task 2009 into an annotation that complies with all of the principles sketched above for deep statistical generation and trained (Bohnet et al., 2010)'s generator on this new annotation.[11] For our experiments, we used the usual training, development and test data split of the WSJ corpus (Langkilde-Geary, 2002; Ringger et al., 2004; Bohnet et al., 2010); Table 1 provides an overview of the used data.

set	section	# sentences
training	2 - 21	39218
development	24	1334
test	23	2400

Table 1: Data split of the used data in the WSJ Corpus

The resulting BLEU score of our experiment was 0.64, which is comparable with the accuracy reported in (Bohnet et al., 2010) (namely, 0.659), who used an annotation that still contained all functional nodes (such that their generation task was considerably more syntactic and thus more straightforward).

To assess furthermore whether the automatically converted PropBank already offers some advantages to other applications than generation, we used it in a semantic role labeling (SRL) experiment with (Björkelund et al., 2010)'s parser. The achieved overall accuracy is 0.818, with all analysis stages (including the predicate identification stage) being automatic, which is a rather competitive figure. In the original CoNLL SRL setting with Oracle reading, an accuracy of 0.856 is achieved.

Another telling comparison can be made between the outcomes of the First Surface Realization Shared Task (Belz et al., 2011), in which two different input representations were given to the competing teams: a shallow representation and a deep representation. The shallow structures were unordered syntactic dependency trees, with all the tokens of the sentence, and the deep structures were predicate-argument graphs with some nodes removed (see Section 2). Although the performance of shallow generators was higher than the performance of the deep generators (the StuMaBa shallow generator (Bohnet et al., 2011a) obtained a BLEU score of 0.89, as opposed to 0.79 of the StuMaBa deep gen-

[11] Obviously, our conversion can be viewed only preliminary. It does not take into account all the subtleties that need to be taken account—for instance, with respect to the information structure; see also Section 6.

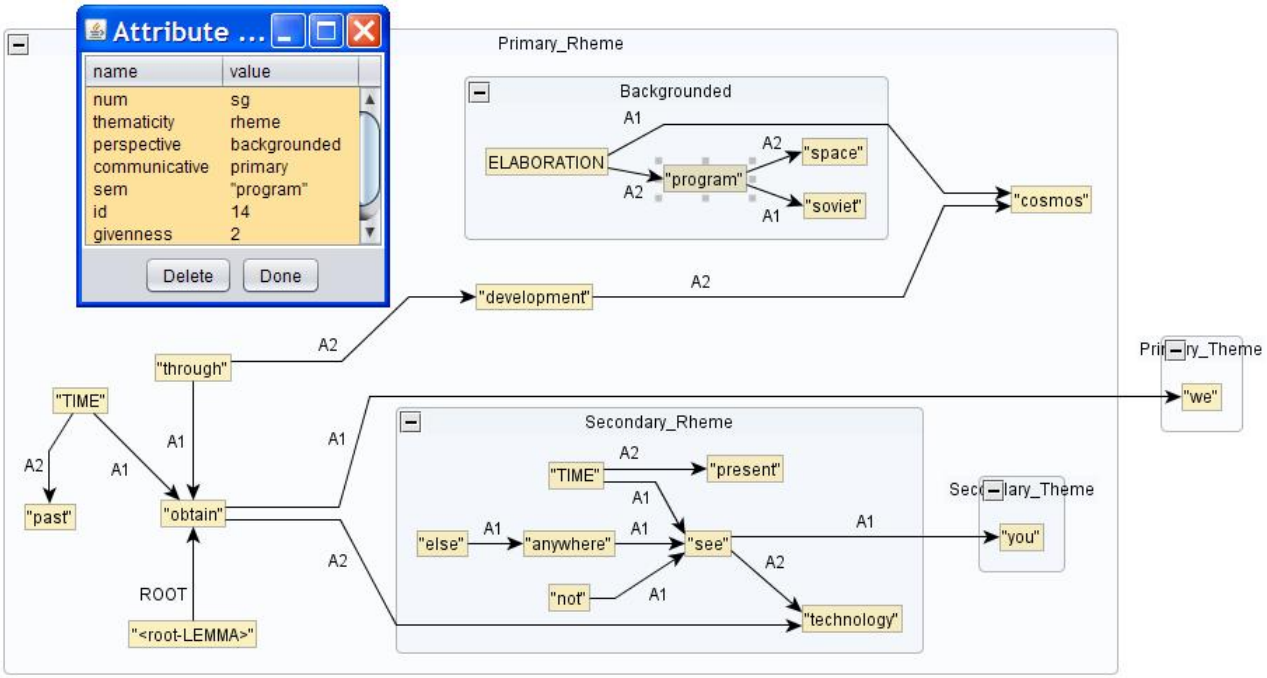

Figure 2: Illustration of the semantic annotation of the sentence *Through the development of Cosmos, the Soviet space program, we obtained technologies you do not see anywhere else.*

erator), the difference is not as striking as one would expect.[12]

6 Conclusions

Our experiments and the Surface Realization Shared Task 2011 suggest that making the deep annotation more semantic does not necessarily imply an unsurmountable problem for stochastic generation. We can thus conclude that deriving automatically a deep semantic annotation from PropBank allowed us to obtain very promising results, both for NLG and SRL. By sticking to universal predicate-argument structures, as PropBank does, we maintain the potential of the corpus to be mapped to other, more idiosyncratic, annotations. Still, automatic conversion will always remain deficient. Thus, a flawless identification of semantic predication cannot be guaranteed. For instance, when an actancial arc points to a preposition, it is not clear how to deduce whether this preposition is semantic or lexical. Also, the treatment of phraseological nodes is problematic, as is the annotation of a comprehensive information structure: the criteria for the automatic derivation of the information structure from the syntactic structure and the topology of a sentence can only be superficial and likely to be even less efficient in longer and complex sentences. The annotation of intersentential coreferences and the identification of gapped elements are further major hurdles for an automatic derivation of a truly semantic resource. As a consequence, we believe that new annotation policies are needed to obtain a high quality semantic resource. The best strategy is to start with a conversion of an existing semantically annotated treebank such as PropBank, revising and extending the result of this conversion in a manual concerted action—always following truly semantic annotation policies.

[12]Note that our results mentioned above cannot be directly compared with the StuMaBa results during the Generation Challenges 2011 because the realizers are different.

Acknowledgments

We would like to thank the reviewers for their valuable comments and suggestions and the Penn Treebank/PropBank/NomBank team, without whom our experiments would not be possible. Many thanks also to Mike White for the useful discussions on some of the topics discussed in the paper. Although we might still not agree on all of the details, he made us see the task of generation-oriented annota-

tion from another angle and revise some of our initial assumptions.

References

S. Bangalore and O. Rambow. 2000. Exploiting a Probabilistic Hierarchical Model for Generation. In *Proc. of COLING '00*.

A. Belz, M. White, D. Espinosa, D. Hogan, and A. Stent. 2011. The First Surface Realization Shared Task: Overview and Evaluation Results. In *ENLG11*.

A. Björkelund, B. Bohnet, L. Hafdell, and P. Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proc. of COLING '10: Demonstration Volume*.

B. Bohnet, L. Wanner, S. Mille, and A. Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proc. of COLING '10*.

B. Bohnet, S. Mille, B. Favre, and L. Wanner. 2011a. <STUMABA>: From Deep Representation to Surface. In *ENLG11*.

B. Bohnet, S. Mille, and L. Wanner. 2011b. Statistical language generation from semantic structures. In *Proc. of International Conference on Dependency Linguistics*.

M. Buch-Kromann, M. Gylling-Jørgensen, L. Jelbech-Knudsen, I. Korzen, and H. Müller. 2011. The inventory of linguistic relations used in the Copenhagen Dependency Treebanks. www.cbs.dk/ content/download/149771/1973272/file.

B. Comrie. 1976. *Aspect*. Cambridge University Press, Cambridge.

P. Cook and J. Payne. 2006. Information Structure and Scope in German. In *LFG06*.

A. Copestake, D. Flickinger, and I. Sag. 1997. Minimal recursion semantics. Technical report, CSLI, Stanford University, Stanford.

J. Gundel, N. Hedberg, and 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.

J.K. Gundel. 1988. "Universals of Topic-Comment Structure". In M. Hammond, E. Moravčik, and J. Wirth, editors, *Studies in Syntactic Typology*. John Benjamins, Amsterdam & Philadelphia.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, and Z. Žabokrtský. 2006. Prague Dependency Treebank 2.0.

K. Knight and V. Hatzivassiloglou. 1995. Two-level, many paths generation. In *Proc. of ACL '95*.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. of COLING/ACL '98*.

I. Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proc. of 2nd INLG Conference*.

F. Mairesse, M Gašić, F. Juričíč, S Keizer, B. Thomson, K. Yu, and S. Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proc. of ACL '10*.

I.A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany.

I.A. Mel'čuk. 2001. *Communicative Organization in Natural Language (The Semantic-Communicative Structure of Sentences)*. Benjamins Academic Publishers, Amsterdam.

M. Mikulová et al. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank: Reference book. www.cbs.dk/content/download/149771/ 1973272/file.

A.H. Oh and A.I. Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *Proc. of ANL/NAACL Workshop on Conversational Systems*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

E. Ringger, M. Gamon, R.C. Moore, D. Rojas, M. Smets, and S. Corston-Oliver. 2004. Linguistically informed statistical models of constituent structure for ordering in sentence realization. In *Proceedings of COLING*, pages 673–679.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel Publishing Company, Dordrecht.

J. F. Sowa. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, USA.

A. Stent, R. Prasad, and M. Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proc. of ACL '04*.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th CoNLL-2008*.

M.A. Walker, O.C. Rambow, and M. Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433.

Y.W. Wong and R.J. Mooney. 2007. Generation by inverting a semantic parser that uses statistical machine translation. In *Proc. of the HLT Conference*.