Grammar Extraction from an HPSG Corpus*

Kiril Iv. Simov
BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
kivs@bgcict.acad.bg

1 Introduction

This paper describes an approach towards extracting an HPSG grammar from an "ideal" HPSG corpus. Our approach follows the methodology for grammar learning from corpora presented in (Bod 1998). His grammar learning model from corpora includes the following elements:

1. Defining the grammar formalism for the target grammar;

2. Establishing a procedure for the construction sentences analyses in the chosen grammar formalism;

3. Establishing a decomposition procedure, which extracts a grammar in the target grammar formalism from the structures in the corpus;

4. A performance model guiding the analysis of new sentences.

Two additional unspoken assumptions are:

5. The structures in the corpus are decomposable into the grammar formalism;

6. The extracted grammar should neither overgenerate, nor undergenerate with respect to the training corpus.

Following this methodology in our work we define a common representation for HPSG corpus and HPSG grammar. Afterwards we define a mechanism for extraction of HPSG grammars from the corpus.

We start with a definition of an "ideal" corpus in general. Such ideal corpus has to ensure the above requirements and assumptions. An "ideal" corpus C in a given grammatical formalism G is a sequence of sentences where each sentence is a member of the set of structures defined as a **strong generative capacity** (SGC) of a grammar Γ in this grammatical formalism: $\forall S.S \in C \rightarrow S \in \mathtt{SGC}(\Gamma)$, where Γ is a grammar in the formalism G. Of course, the grammar Γ is unknown. It is implicitly represented in the corpus C.

We are using this general definition in order to define an "ideal" corpus in HPSG. Thus we choose:

1. A logical formalism for HPSG — King's Logic (SRL) (King 1989);

2. A definition of strong generative capacity in HPSG as a set of feature structures closely related to the special interpretation in SRL (exhaustive models) along the lines of (King 1999) and (Pollard 1999).

3. A corpus in HPSG will contain sentences that are members of $\mathtt{SGC}(\Gamma)$ for some grammar Γ in SRL.

In this paper we don't define a performance model for the extracted grammar.

The structure of the paper is as follows: first we present a formalism for HPSG as feature graphs based on SRL, then we define the extraction of feature graphs from an HPSG corpus and how we can construct HPSG grammar on the base of the extracted graphs.

2 Formalism for HPSG

In this section we present a logical formalism for HPSG. Then a normal form (exclusive matrices) for finite theory in this formalism is defined as a set of feature graphs. These graphs are considered as a representation of grammars and corpora in HPSG. First, we present the syntax of King's Logic. For

* This work is funded by the Volkswagen Stiftung, Federal Republic of Germany under the Programme "Cooperation with Natural and Engineering Scientists in Central and Eastern Europe" contract I/76 887.

full description of it see (King 1989).

2.1 King's Logic — SRL

$\Sigma = \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ is a finite **SRL signature** iff $\mathcal{S}$ is a finite set of *species*, $\mathcal{F}$ is a set of *features*, and $\mathcal{A} : \mathcal{S} \times \mathcal{F} \rightarrow Pow(\mathcal{S})$ is an *appropriateness function.*

τ is a term iff τ is a member of the smallest set $\mathcal{T}$ such that (1) $: \in \mathcal{T}$, and (2) for each $\phi \in \mathcal{F}$ and each $\tau \in \mathcal{T}$, $\tau\phi \in \mathcal{T}$.

δ is a **description** iff δ is a member of the smallest set $\mathcal{D}$ such that (1) for each $\sigma \in \mathcal{S}$ and for each $\tau \in \mathcal{T}$, $\tau \sim \sigma \in \mathcal{D}$, (2) for each $\tau_1 \in \mathcal{T}$ and $\tau_2 \in \mathcal{T}$, $\tau_1 \approx \tau_2 \in \mathcal{D}$ and $\tau_1 \not\approx \tau_2 \in \mathcal{D}$, (3) for each $\delta \in \mathcal{D}$, $\neg\delta \in \mathcal{D}$, (4) for each $\delta_1 \in \mathcal{D}$ and $\delta_2 \in \mathcal{D}$, $[\delta_1 \wedge \delta_2] \in \mathcal{D}$, $[\delta_1 \vee \delta_2] \in \mathcal{D}$, and $[\delta_1 \rightarrow \delta_2] \in \mathcal{D}$. Each subset $\theta \subseteq \mathcal{D}$ is an **SRL theory**. (King 1989) defines a standart model theoretical semantics for his logic.

An HPSG grammar $\Gamma = \langle \Sigma, \theta \rangle$ in SRL consists of: (1) a signature Σ which gives the ontology of entities that exist in the universe and the appropriateness conditions on them, and (2) a theory θ which gives the restrictions upon these entities. Usually the descriptions in the theory part are implications.

(King 1999) and (Pollard 1999) define their notions of strong generative capacity in HPSG based on SRL as logical formalism. We are using an approximation of their definitions of `SGC` based on a normal form in SRL, called **exclusive matrix** (see (King and Simov 1998)). We present the elements of the normal form as feature graphs. One important point about our feature graphs is that they are considered as descriptions in SRL and thus syntactic entities but not as semantic entities — elements of an interpretation.

2.2 Feature Graphs

Let $\Sigma = \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ be a finite signature. A **feature graph** with respect to Σ is a directed, connected and rooted graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{V}, \rho, \mathcal{S} \rangle$ such that: (1) $\mathcal{N}$ is a set of **nodes**, (2) $\mathcal{V} : \mathcal{N} \times \mathcal{F} \rightarrow \mathcal{N}$ is a partial **arc function**, (3) ρ is a **root node**, (4) $\mathcal{S} : \mathcal{N} \rightarrow \mathcal{S}$ is a total **species assignment function**, and (5) for each $\nu_1, \nu_2 \in \mathcal{N}$ and each $\phi \in \mathcal{F}$ such that $\mathcal{V}\langle \nu_1, \phi \rangle \downarrow$ and $\mathcal{V}\langle \nu_1, \phi \rangle = \nu_2$, then $\mathcal{S}\langle \nu_2 \rangle \in \mathcal{A}\langle \mathcal{S}\langle \nu_1 \rangle, \phi \rangle$. We say that the feature graph $\mathcal{G}$ is **finite** if and only if the set of nodes is finite. A feature graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{V}, \rho, \mathcal{S} \rangle$ such that for each node $\nu \in \mathcal{N}$ and each feature $\phi \in \mathcal{F}$ if $\mathcal{A}\langle \mathcal{S}\langle \nu \rangle, \phi \rangle \downarrow$ then $\mathcal{V}\langle \nu, \phi \rangle \downarrow$ is called a **complete feature graph**. For each graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{V}, \rho, \mathcal{S} \rangle$ and node ν in $\mathcal{G}$ with $\mathcal{G} \mid_\nu = \langle \mathcal{N}_\nu, \mathcal{V} \mid_{\mathcal{N}_\nu}, \nu, \mathcal{S} \mid_{\mathcal{N}_\nu} \rangle$ we denote the **subgraph** of $\mathcal{G}$ starting on node ν.

For each two graphs $\mathcal{G}_1 = \langle \mathcal{N}_1, \mathcal{V}_1, \rho_1, \mathcal{S}_1 \rangle$ and $\mathcal{G}_2 = \langle \mathcal{N}_2, \mathcal{V}_2, \rho_2, \mathcal{S}_2 \rangle$ we say that graph $\mathcal{G}_1$ **subsumes** graph $\mathcal{G}_2$ ($\mathcal{G}_2 \sqsubseteq \mathcal{G}_1$) iff there is an *isomorphism* $\gamma : \mathcal{N}_1 \rightarrow \mathcal{N}_2'$, $\mathcal{N}_2' \subseteq \mathcal{N}_2$, such that (1) $\gamma(\rho_1) = \rho_2$, (2) for each $\nu, \nu' \in \mathcal{N}_1$ and each feature ϕ, $\mathcal{V}_1\langle \nu, \phi \rangle = \nu'$ iff $\mathcal{V}_2\langle \gamma(\nu), \phi \rangle = \gamma(\nu')$, and (3) for each $\nu \in \mathcal{N}_1$, $\mathcal{S}_1\langle \nu \rangle = \mathcal{S}_2\langle \gamma(\nu) \rangle$. For each two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ if $\mathcal{G}_2 \sqsubseteq \mathcal{G}_1$ and $\mathcal{G}_1 \sqsubseteq \mathcal{G}_2$ we say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are **equivalent.**

For finite feature graphs we could define a translation to a SRL descriptions using the correspondences between paths in the graph and terms. Thus we can interpret each finite feature graph as a description in SRL. Using the set of all finite feature graphs that subsume a given infinite feature graph we can define also interpretation of each infinite feature graph. Moreover, we can define a correspondence between exclusive matrices and feature graphs. Thus using the algorithm from (King and Simov 1998) and adding the information from the signature as a special theory[1] we can represent each finite SRL theory as a set of feature graphs.

Thus, feature graphs can be used for both: (1) **Representation of an HPSG corpus.** Each sentence in the corpus is represented as a complete feature graph. One can easily establish a correspondence between the elements of strong generative capacity of an HPSG grammar and complete feature graphs. Thus complete feature graphs are a good representation for an HPSG corpus, and (2) **Representation of the extracted grammar as a set of feature graphs.** The construction of a graph representation of a finite theory demonstrates that using fea-

[1] In order to account for the information in the signature we construct a special theory

$$\theta_\Sigma = \{ \bigvee_{\sigma \in \mathcal{S}} [\bigwedge_{\mathcal{A}(\sigma,\phi) \neq \emptyset, \phi \in \mathcal{F}} [: \phi \approx : \phi]] \}.$$

Then for each theory θ we form the theory $\theta^e = \theta \cup \theta_\Sigma$ which is semantically equivalent to the original theory.

ture graphs as grammar representation doesn't impose any restrictions over the class of possible finite grammars in SRL.

3 Extracting Grammars

In this section we present a fragmentation of an HPSG corpus into a set of feature graphs such that its subsets with particular properties can comprise HPSG grammars.

Let C be an HPSG corpus comprising a set of complete feature graphs. Let G be a grammar represented as a set of feature graphs as described above. We say that G is a grammar of the corpus C if and only if for each graph $\mathcal{G}_C$ in C and each node $\nu \in \mathcal{G}_C$ there is a graph $\mathcal{G}_G$ in G such that $\mathcal{G}_C \mid \nu \sqsubseteq \mathcal{G}_G$. If G is a grammar of C then we could construct a model of G by the graphs in C.

Now we define the set of fragments extracted from the corpus C. We construct a set FGF of feature graphs such that

1. For each graph $\mathcal{G} \in FGF$, $\mathcal{V}\langle\rho,\phi\rangle \downarrow$ iff $\mathcal{A}\langle\mathcal{S}\langle\rho\rangle,\phi\rangle\downarrow$, and
2. For each graph $\mathcal{G} \in FGF$, there is a graph $\mathcal{G}_C$ in C and there is a node $\nu \in \mathcal{G}_C$ such that $\mathcal{G}_C \mid \nu \sqsubseteq \mathcal{G}$.

The first condition ensures that all features appropriate for a given species will be presented at the root node for each feature graph in FGF whose root is labelled by this feature. The second condition ensures that each feature graph in FGF is really a fragment of at least one feature graph in the corpus. FGF contains feature graphs with different size. We ordered FGF according to the subsumption relation over feature graphs. The set ordered in this way is a set of partial orders over the features graphs in FGF.

Let G be a set of feature graphs such that for each minimal feature graph M in FGF there is at least one feature graph in G that subsumes M and G contains only feature graphs from FGF. Each grammar G constructed in this way is a grammar of corpus C. Choosing different feature graphs in FGF we can construct different grammars. In order to extract the best grammar we will need to impose some external requirements. Such requirements can be that the extracted grammar ensures the shortest inferences for the sentences in the corpus or that it contains the most frequent graphs in the corpus.

An alternative to the grammar construction as it is described above is to use the whole set FGF as a grammar. In fact, this is the approach taken by Rens Bod in his work ((Bod 1998)) with respect to other grammar formalisms.

4 Conclusion and Future Work

In the paper we define the notion of "ideal" corpus in HPSG. We develop a common representation of such kind of corpus and the grammars that can be extracted from it. Then we present an approach for grammars extraction from such a corpus.

There are still open questions. As the set FGF could be very large it is necessary to develop mechanisms for compact representation of it or an algorithm for extracting a concrete grammar with desirable properties. If one would like to use the whole set FGF as a grammar then a procedure for construction of sentence analyses will be necessary.

References

(Bod 98) Rens Bod. *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications, CSLI, California, USA, 1998.

(King 1989) Paul J. King. *A Logical Formalism for Head-Driven Phrase Structure Grammar*. Doctoral thesis, Manchester University, Manchester, England. 1989.

(King 1999) Paul J. King. *Towards Thruth in Head-Driven Phrase Structure Grammar*. In V. Kordoni (Ed.), *Tübingen Studies in HPSG,* Number 132 in Arbeitspapiere des SFB 340, pp 301-352. Germany. 1999.

(King & Simov 1998) Paul J. King and Kiril Iv. Simov. The automatic deduction of classificatory systems from linguistic theories. In *Grammars*, volume 1, number 2, pages 103-153. Kluwer Academic Publishers, The Netherlands. 1998.

(Pollard 1999) Carl Pollard. *Strong Generative Capacity in HPSG*. in Webelhuth, G., Koenig, J.-P., and Kathol, A., editors, *Lexical and Constructional Aspect of Linguistic Explanation,* pp 281-297. CSLI, Stanford, California, USA. 1999.