# Logical investigations on the adequacy of certain feature-based theories of natural language

**Anders Søgaard**
Center for Language Technology
Njalsgade 80
DK-2300 Copenhagen
anders@cst.dk

## Abstract

A theory of natural language can be evaluated on both extensional and intensional grounds. Systematic investigations of the extension of a theory may, for instance, lead to studies of the invariance properties of such theories. The intentional parameters that I wish to address include complexity, learnability, and monotonicity. The main results, on which my thesis builds, up to this point, include: (i) the universal recognition problem of model-theoretic feature-based grammar formalisms is complete for non-deterministic polynomial time, since such formalisms have the polysize model property, (ii) this result holds also for linearization-based extensions, (iii) the universal recognition problem of strongly monotonic, hybrid feature-based grammar formalisms is decidable in deterministic polynomial time, and (iv) there exists a strongly monotonic unification categorial grammar that is learnable in the limit from positive data. In addition, invariance studies have lead to the identification of a class of modal languages that define common feature-based grammar formalisms. The objective of my studies is to identify a tractable and learnable feature-based formalism.

## 1 Introduction

My work addresses certain extensional and intensional properties of various feature-based theories of natural language, incl. unification categorial grammar (Zeevat, 1988) and head-driven phrase structure grammar (Pollard and Sag, 1994). The theories are referred to henceforth as UCG and HPSG. A feature-based theory of natural language defines a set of feature-based grammars (and interfaces). A feature-based grammar associates feature structures with the strings of the language in question. Consequently, it makes sense to start off with a definition of a feature structure. A signature $\langle \mathsf{Lbls}, \mathsf{Atmc} \rangle$ is a pair of sets of labels and atomic informations. In UCG and HPSG, both are finite. A feature structure of a signature $\langle \mathsf{Lbls}, \mathsf{Atmc} \rangle$ is then an ordered triple $\langle \mathbb{N}, \{R_\lambda\}_{\lambda \in \mathsf{Lbls}}, \{Q_\alpha\}_{\alpha \in \mathsf{Atmc}} \rangle$, where $\mathbb{N}$ is a set of nodes, $R_\lambda$ is a partial function, and $Q_\alpha$ is a unary one, for all $\lambda \in \mathsf{Lbls}$ and $\alpha \in \mathsf{Atmc}$.

Grammars employ feature structures in different ways. A *hybrid* grammar, in its most raw format, is a 4-tuple $\langle \langle \mathsf{Lbls}, \mathsf{Atmc} \rangle, \mathbb{V}, \mathsf{Rules}, \mathtt{start} \rangle$, where $\mathbb{V}$ is the vocabulary. One may add a specification function such that, for instance, $\forall x \, \exists y \; R_\lambda(x, y) \rightarrow Q_\alpha(y)$. Intuitively, if $\mathcal{L}(\mathcal{G})$ is the language of $\mathcal{G}$, and if $\mathcal{G}$ is a hybrid grammar, $\mathcal{L}(\mathcal{G})$ is the set of strings "modelled" (derivable) by the grammar, whereas the language of a model-theoretic grammar is the set of strings that (or whose relational structures) model the grammar. The generative-enumerative core of a hybrid grammar is in its set of rules ($\mathsf{Rules}$).

*Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 239–242,
New York, June 2006. ©2006 Association for Computational Linguistics

The language of $\mathcal{G}$ is defined as

$$\mathcal{L}(\mathcal{G}) = \{x \in \mathbb{V}^* | \exists f \in \mathcal{F}\ \texttt{start} \sqsubseteq f \wedge f \Rightarrow x\}$$

where $\mathcal{F}$ is the set of feature structures, and $f' \sqsubseteq f$ means that the information in $f'$ is also in $f$, and $f \wedge f'$ is consistent. Finally, $f \Rightarrow \sigma$ means that $\sigma$ is derivable from $f$ by $\mathsf{Rules}$. It should be obvious that a hybrid grammar is "hybrid" in the sense that it combines generative rules and subsumption ($\sqsubseteq$), which is essentially model-theoretically defined, i.e. $f' \sqsubseteq f$ iff if $f \models \phi$ then $f' \models \phi$.

Consider a grammar $\mathcal{G}'$, which is entirely model-theoretic, i.e. the grammar is axiomatically defined, and feature structures are seen as Kripke frames. In other words, a *model-theoretic* grammar is a 4-tuple $\langle\langle \mathsf{Lbls}, \mathsf{Atmc}\rangle, \mathbb{V}, \mathsf{Axms}, \texttt{root}\rangle$, where rules have been replaced with a set of axioms $\mathsf{Axms}$ defined in some logic, and the $\texttt{start}$ category is replaced with a $\texttt{root}$ proposition, defined in the axioms. The signature is now a signature of modalities and propositions. It is important, in order to maintain the overall picture, to remember that on the standard translation into first order logic, modalities and propositions translate into binary and unary relations, respectively. Consequently, this is, at the moment, just a notational change. The introduction of modal vocabulary is relevant to the specification of feature-based theories later on.

The universal recognition problem amounts to this question: Given some pair $\sigma, \mathcal{G}$, $\sigma \in \mathcal{L}(\mathcal{G})$? In particular, when it is said that the universal recognition problem of some formalism is in some complexity class, it means that there exists an algorithm such that the membership of any string in any grammar licensed by the formalism can be decided in the time complexity of that class by running the algorithm. The universal recognition problems of model-theoretic UCG and HPSG, and the linearization-based extension of the latter, and strongly monotonic HPSG are examined in a minute.

Our introductions of UCG and HPSG are of course only partial, since this paper is of limited length. In fact, no more than a paragraph is spend on these introductions:

**Unification categorial grammar** UCG and HPSG are both said to be sign-based, i.e. the fundamental unit is the sign. A sign in UCG has the structure W:C:S:O, where W contains information about the phonology of the sign, C presents its syntactic category, S is the semantics, and O constrains word order in determining how the sign combines with other signs. Signs combine by functional application (instantiation and stripping). Instantiation checks if the active part of the syntactic category of the functor unifies with the syntactic category of the argument, and if unification succeeds, the instantiated functor is stripped, and the phonology features are concatenated. Type hierarchies extend UCG in a natural way. Instantiation and stripping can be interpreted as phrasal types rather than functions. Model-theoretic parsing of some string $\sigma \in \mathcal{L}(\mathcal{G})$ then amounts to finding a connected and rooted (minimal) model $\mathcal{M}$ whose linearization is $\sigma$, s.t. $\mathcal{M}, w \in [\![\texttt{root}]\!] \models \texttt{Axms}$.

**Head-driven phrase structure grammar** HPSG parsing is much the same, except $\mathsf{Axms}$ is conjoined with $\mathsf{Prncp}$, the set of linguistic principles. One traditional problem with HPSG is that it employs sets. Some recent (computationally oriented) versions of HPSG substitute sets with so-called "diff-lists", which are briefly lists with pointers to their last elements, and for now we settle with this option. An alternative is mentioned in our discussion of linearization-based HPSG, namely a simulation of sets as underspecified lists; or one can perhaps employ polyadic modalities ($n$-ary relations). The linguistic principles in $\mathsf{Prncp}$ include, for instance, the head feature principle, which says that in a headed phrase, the HEAD value of the mother is identical to that of the head daughter, the immediate dominance principle and the weak coordination principle.

## 2 Some formal results

Our first complexity result, i.e. (i) the universal recognition problem of model-theoretic feature-based grammar formalisms is complete for nondeterministic polynomial time, since such formalisms have the polysize model property, is obtained by specification of UCG and HPSG in some modal language that has a model checking problem of polynomial time complexity. The model checking amounts to evaluating a formula $\phi$ in a model $\mathcal{M}$. If a formalism has the polysize model property, its universal recognition problem can be evaluated on small models that are polynomial in the size of the strings. If the specification language has a polynomial model checking problem, a model can thus be non-deterministically chosen and evaluated in polynomial time, and the result follows. Consequently, the quest is two-fold: It is necessary to establish the polysize model property for UCG and HSPG, and we then need to identify an adequate specification language that embeds these theories. The polysize model property follows from Lemma 2.1.[1]

**Lemma 2.1.** *Say $\phi$ represents a* UCG *or* HPSG *recognition problem for a string $\sigma$. If there exists a model $\mathcal{M}$ and a node $w \in \mathbb{N}$ s.t. $\mathcal{M}, w \in [\![\texttt{root}]\!] \models \phi$, then there also exists a model $\mathcal{M}'$ of at most k cardinality and a node $w' \in \mathbb{N}$ s.t. $\mathcal{M}', w' \in [\![\texttt{root}]\!] \models \phi$, where $k = (2|\sigma| - 1) \times (u + 1) \times$* **paths***, where u is the number of unary rules in the grammar, and* **paths** *is a constant that depends on the non-recursive part of the feature geometries of* UCG *and* HPSG*. In particular,* **paths** $= |\{\pi \in \texttt{Lbls}^* |$ *no label occurs twice in* $\pi\}|$.

It is now left to show that UCG and HPSG can be specified (defined) in some formal language that has a polynomial time model checking problem. Since UCG subsumes HPSG, it suffices to show that this holds for HPSG. Various translations of HPSG into specification languages have been proposed, and my recent work includes a couple of such translations, but in this synopsis, to save space, we refer to the specification language of Kracht (1995). He defines a translation of HPSG into $\mathsf{PDL}^{\cup,[*]}$, propositional dynamic logic with intersection and the master modality. The master modality is defined s.t. $\mathcal{M}, w \models [*]\phi$ iff $\forall w'((w, w') \in (\bigcup_{\alpha \text{ is atomic}} R_\alpha) \& \mathcal{M}, w' \models \phi)$. It is trivial to show the high undecidability of this language, for instance, the recurrent tiling problem can be encoded in $\mathsf{PDL}^{\cup,[*]}$. The model checking of $\mathsf{PDL}^{\cup,[*]}$ is indeed decidable in polynomial time; this is evident from the investigations of Lange (2006). Consequently, Theorem 2.2 follows.

**Theorem 2.2.** *The universal recognition problem of* UCG *and* HPSG *is decidable in non-deterministic polynomial time.*

The result can be extended to linearization-based versions of model-theoretic UCG and HPSG. On the model-theoretic perspective, immediate dominance and linear precedence are already split, since they are represented by different modalities. The thing to do when languages of freer word-order are considered, is then simply to relax the linearization of immediate dominance principles. The master modality of $\mathsf{PDL}^{\cup,[*]}$ can be used to implement weak linear precedence. Weak linear precedence is thus, in some sense, constraints on an underspecified list of strings, and domain union, for instance, is "unification" of underspecified lists.

Strong monotonicity has been mentioned a couple of times. The notion is relevant on a hybrid set-up. Some grammar formalisms are non-monotonic in the traditional sense, but we confine ourselves to monotonic ones, for the simple reason that the modal languages considered here are all monotonic. The notion of *strong* monotonicity is different. Consider a conventional context-free grammar. On our definition of strong monotonicity, a context-free grammar $\mathcal{G}$ of $\mathcal{L}$ is *not* strongly monotonic if it is ambiguous on $\mathcal{L}$, i.e. if there exists a string $\sigma \in \mathcal{L}$, such that more than one tree can be derived by $\mathsf{Rules}_\mathcal{G}$. The strong monotonicity hypothesis, i.e. that natural language grammars are strongly monotonic, is very strong and somewhat unnatural to most linguists. Since Linguistics 101, we were taught that languages are inherently ambiguous. Strongly monotonic grammars of course have formal interest, since they exhibit a number of nice properties, discussed in the next paragraph, but they *need* not be irrelevant

[1] Unary rules only apply once to the same unary extension in Lemma 2.1. In the proof of Theorem 2.3, a unary extension is the result of a single application, i.e. $u = 1$ in Lemma 2.1. It is not clear to me what the linguistic relevant restriction is.

in linguistics either. In feature-based grammars that employ type hierarchies, it is possible, after all, to underspecify ambiguities. It has been argued that such underspecification is possible and a linguistically interesting option in the context of both quantification, attachment ambiguities, and the combinatorics of case and word order.

Say $\mathcal{G}$ is a hybrid grammar and strongly monotonic. For one thing, this means that the lexicon in $\mathsf{Rules}_{\mathcal{G}}$ is rigid s.t. a partial function map strings onto feature structures. It also means that $\phi$ has a *unique* model of size less than or equal to $k$. A rather restrictive parsing algorithm is introduced: Say $\mathsf{Rules}_{\mathcal{G}}$ consists of $b$ binary rules and $u$ unary ones. $\mathcal{G}$ tries to combine pairs of constituents bottom-up by $b$, and if this does not succeed, $u$ is used to extend any of the constituents by a single application. On the assumption that $\mathsf{Rules}$ contains no unary rules, $(\frac{|\sigma|^2-|\sigma|}{2}) \times b$ is the number of possible projections. When unary rules are added, this number is multiplied by the number of unary rules times the number of binary rules, since the binary rules are first tried out, and if that doesn't work, unary rules are used to extend nodes, and binary rules are applied again. The algorithm only has to run once because of strong monotonicity. Consequently, Theorem 2.3 holds:

**Theorem 2.3.** *The universal recognition problem of strongly monotonic and hybrid feature-based grammars is decidable in deterministic polynomial time.*

*Proof.* $(\frac{|\sigma|^2-|\sigma|}{2}) \times b$ is the number of possible projections in the abscence of unary rules. Add unary rules and the number of possible projections is

$$\frac{b\times(|\sigma|^2-|\sigma|)}{2}(u+1)^3+|\sigma|u$$

For each step, unification is tested. Unification is decidable in time $\Theta(\delta \times \omega(\delta))$ (Hegner, 1991), where $\delta$ is the number of distinct edges in the two feature structures, i.e. $\delta = \textbf{paths}$ in the above, and $\omega(\delta)$ is the inverse Ackermann function. For all practical purposes, $\omega(\delta)$ is lower than 5 (Hegner, 1991). Nothing else has to be computed to decide universal recognition for a strongly monotonic hybrid feature-based grammar. The result follows. □

The learnability result, "(iv)" in the above, derives from a result established by Kanazawa (1998), namely that rigid categorial grammars are leanable in the limit, even from positive data. If so it follows that there exists strongly monotonic unification categorial grammars that are also learnable in the limit from positive data, since strongly monotonic grammars are rigid, by definition, and since simple unification categorial grammars can be embedded in classical ones.

I envisage a tractable and learnable feature-based grammar formalism to look much like strongly monotonic UCG extended with type hiearchies and linearization. The notion of strong monotonic can be relativized in various ways without loosing tractability, and this line of research should be pursued.

## References

Stephen Hegner. 1991. Horn-extended feature structures. In *The 5th European Chapter of the Association for Computational Linguistics*, pages 33–38, Berlin, Germany.

Makoto Kanazawa. 1998. *Learnable classes of categorial grammars.* CSLI Publications, Stanford, California.

Marcus Kracht. 1995. Is there a genuine modal perspective on feature structures? *Linguistics & Philosophy*, 18:401–458.

Martin Lange. 2006. Model checking propositional dynamic logic with all extras. *Journal of Applied Logic*, 4:39–49.

Carl Pollard and Ivan Sag. 1994. *Head-driven phrase structure grammar*, volume 4 of *Studies in Contemporary Linguistics.* The University of Chicago Press, Chicago, Illinois.

Henk Zeevat. 1988. Combining categorial grammar and unification. In Uwe Reyle and Christian Rohrer, editors, *Natural language parsing and linguistic theories*, pages 202–229. Reidel, Dordrecht, Germany.