HPSG-DOP: Towards Exemplar-based HPSG

Doug Arnold & Evita Lindardaki
Dept. of Language & Linguistics
University of Essex
Colchester, UK
doug,elinaro@essex.ac.uk

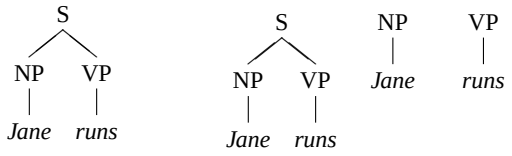
Abstract

Data Oriented Parsing (DOP) is an exemplar-based model of language use that processes new input based on past experience by combining structural fragments extracted from a given treebank. In the simplest case these fragments are subparts of simple phrase structure trees (Tree-DOP), each associated with some probability. The approach is attractive in many ways but the impoverished representational basis is a serious drawback from a linguistic point of view. This paper describes the theoretical characteristics of a novel linguistically richer version of DOP based on the Head-driven Phrase Structure Grammar (HPSG) formalism.

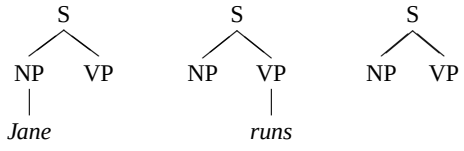
1 Introduction

The evidence of the probabilistic properties displayed in human language processing (Juliano and Tanenhaus, 1993; Jurafsky, 1996) has led to the statistical enrichment of Natural Language Processing (NLP) models. One approach to this involves associating the rules of a competence grammar with probabilities computed from large-scale syntactically annotated corpora. Simply adding probabilities to rules, however, cannot provide an optimal criterion because disambiguation preferences are "memory-based" and can depend on arbitrarily large syntactic constructions (Bod, 2003). Exemplar-based models of language use have gained ground in recent research.

A well-known model based on such an approach is Data Oriented Parsing (DOP) (Bod, 1992; Bod, 1995) which processes new input by combining fragments extracted from a given treebank. In the simplest case these fragments are subparts of simple phrase structure trees (Tree-DOP) produced by two decomposition operations; $Root$ and $Frontier$. $Root$ creates passive fragments by extracting substructures as in Figure 1(*b*) , while $Frontier$ produces active fragments by deleting pieces of substructure as in Figure 1(*c*) . Each fragment is assigned a probability based on some predefined estimator. Disambiguation involves finding the structure(s) with the highest probability.



(*a*) Initial structure. (*b*) The *Root* operation.



(*c*) The *Frontier* operation.

Figure 1: Decomposition of *Jane runs* in Tree-DOP.

The approach is attractive in many ways but the impoverished representational basis is a serious drawback from a linguistic point of view. Bod and Kaplan (1998) address this issue by proposing a linguistically richer version of DOP based on the more

Presented at 'Exemplar-Based Models of Language Acquisition and Use' Workshop, ESSLLI07, Dublin, Aug 2007.

sophisticated Lexical Functional Grammar (LFG) representations. Even though LFG-DOP constitutes a very powerful model of language performance, it also suffers from several disadvantages. The first of these relates to the degree of generality of the produced fragments. The traditional decomposition operations create fragments that are over-specific, leading to under-generation and exacerbating the normal problem of data sparsity. $Root$ and $Frontier$, for example, produce fragments like (1) a and b from the corpus representation of *"Jane runs"*. These fragments, however, will not be useable in parsing either *"Sam likes Jane"*, because *Jane* is not *acc*, or *"Jack runs"*, assuming *Jack* is not *fem*.

(1) a. [$_{NP}$ Jane]$_{3rd/sg/fem/nom}$
b. [$_{S}$ NP$_{3rd/sg/fem/nom}$ runs]

To overcome this, Bod and Kaplan formulate a third decomposition operation known as $Discard$ which generalises over the fragments produced by the other two. $Discard$, however, applies in a highly unconstrained manner causing, on the one hand, the size of the fragment corpus to explode and allowing, on the other, under-specific fragments to be produced (2). The latter leads to overgeneration problems (e.g. *"Him runs"*) which require a corpus-based redefinition of the notion of grammaticality.

(2) [$_{S}$ NP$_{3rd/sg}$ runs]

In addition, not all of LFG's well-formedness conditions can be checked efficiently during the derivation process, resulting in some probability mass being assigned to invalid structures. This probability mass is hence "wasted" raising theoretical questions about the particular disambiguation algorithm.

In the following section we set the theoretical background of a novel version of DOP based on Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), that addresses these issues. Previous attempts to define such a model (Neumann, 1999; Neumann, 2003) were based on extracting a Stochastic Lexicalised Tree Grammar (SLTG) from an HPSG parsed training corpus and using it in a manner similar to Tree-DOP. Node labels in the trees represent the HPSG rule-schema applied during the corresponding derivation step. A complete parse tree can be unfolded into an HPSG representation by expanding the rule labels and lexical types to the corresponding feature structures. Despite their differences, this approach suffers in some cases from the same problems as LFG-DOP. Well-formedness of the resulting structure, for example, cannot be ensured in an efficient manner because some of the parse trees produced cannot be unfolded into valid feature structures. This approach, however, does not provide the only, or even the most linguistically enhanced way of moving towards HPSG-DOP.

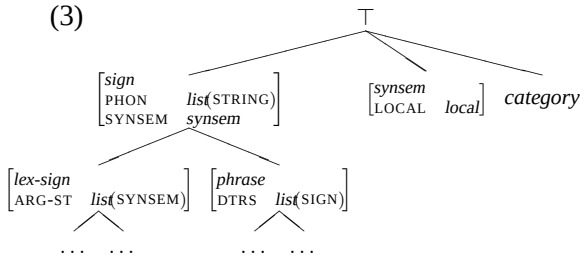
2 HPSG-DOP

Presenting a DOP model involves instantiating the following four parameters: (i) how utterances are represented; (ii) how representations are decomposed into fragments; (iii) how fragments are combined; and (iv) how the proposed analyses are disambiguated.

2.1 Representation

The representational framework we assume is conventional HPSG, along the lines of (Ginzburg and Sag, 2000). The HPSG linguistic ontology is a system of *signs*. These can be either of type *phrase* describing phrasal constituents or *lex-sign* describing words and lexemes (3). All *signs* have the top level attributes PHON and SYNSEM which describe the phonological content and the syntactico-semantic characteristics of the *sign* in question respectively. In addition, *signs* of type *phrase* carry the attribute DTRS (daughters) describing the surface constituency of the phrase. Lexical *signs*, on the other hand, possess the feature ARG-ST whose value is an ordered list of objects corresponding to the arguments (subject, specifier and any complements) required by the *lex-sign* being described.

(3)



We will draw feature structures either as Directed Acyclic Graphs (DAGs) or as Attribute Value Ma-

trices (AVMs), using a wide range of abbreviations (e.g. 'NP' stands for a nominal phrase with empty SPR and COMPS lists). Figure 2 gives the DAG representation of *"Jane runs"*, while Figure 3 presents it as an AVM (both somewhat simplified).

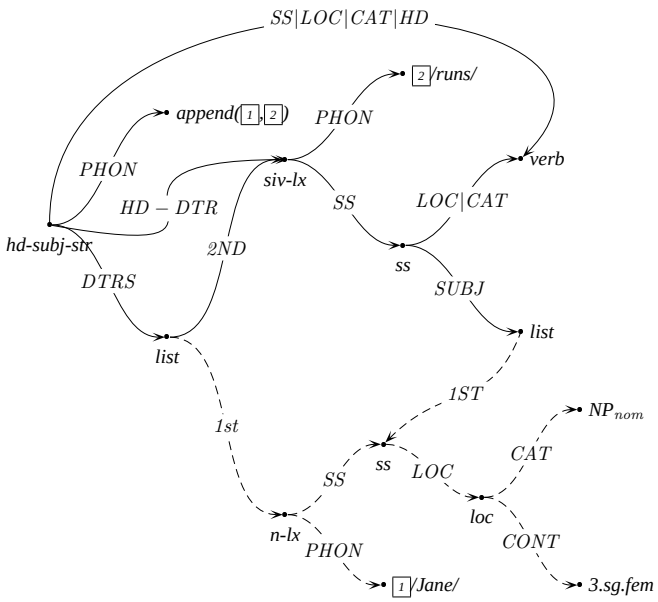


Figure 2: DAG representation of *Jane runs*.

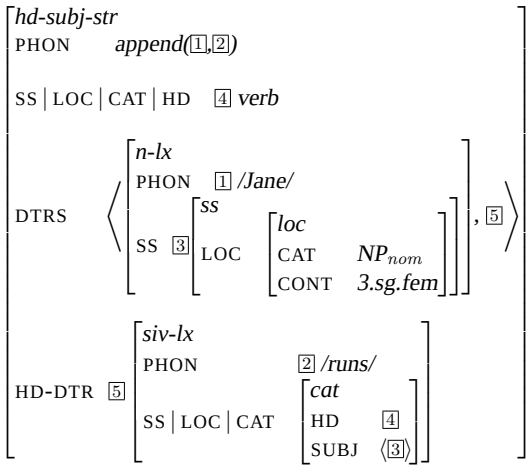


Figure 3: AVM representation of *Jane runs*.

We assume representations are *totally well-typed* feature structures. A representation is *totally well-typed* if all and only the required attributes are present and each of them has an appropriate value. Of course, this only makes sense against the background of a particular *type theory*, that is, a *signature*, which defines an hierarchy of *types*, and a collection of *type constraints* which indicate what combinations of attributes and values are permitted for different types. Fragments should respect the same principles as the representations they are produced from: i.e. they should be *totally well-typed* feature structures. The *total well-typedness* requirement implies that fragments may be subject to a form of type inference which we will refer to as *type expansion*:

Definition 2.1 ($TypeExp$**)** *Let F be a feature structure, and T a type theory, then $TypeExp_T(F)$ is the most general totally well-typed extension of F according to T such that $F \sqsubseteq TypeExp_T(F)$.*

Type expanding the sort *phrase*, for example, produces the feature structure in Figure 4 (assuming the type theory in (Ginzburg and Sag, 2000)).

phrase	
PHON	*list-of-phonemes*
SYNSEM	*synsem*
DTRS	*list-of-signs*
HD-DTR	*sign*

Figure 4: TypeExp(*phrase*)

Type expansion can produce very specific results. In addition to adding feature-value pairs, the constraints on certain types can enforce re-entrancies between various parts of the feature structure (e.g. the type *hd-subj-struc* requires that the SUBJ value of its head-daughter and the SYNSEM value of its non-head daughter are structure-shared).

2.2 Decomposition Operations

Decomposition in HPSG-DOP is carried out by $Root$ and $Frontier$. Before extending these operations so that they become applicable to feature structures, we will introduce some terminology. Let the notion of '*descendants*' (of a *sign*), be recursively defined as the elements of the *sign*'s $DTRS$ list, and their descendants. In addition, let F be a feature structure with a descendant D_F. Suppose D_F is removed from F giving rise to F'. Then $Context\ (D_F)$ denotes the subgraph rooted at the removal node in $TypeExp_T(F')$, and $Inherent\ (D_F)$ denotes the set-theoretic relative complement of D_F and $Context\ (D_F)$.

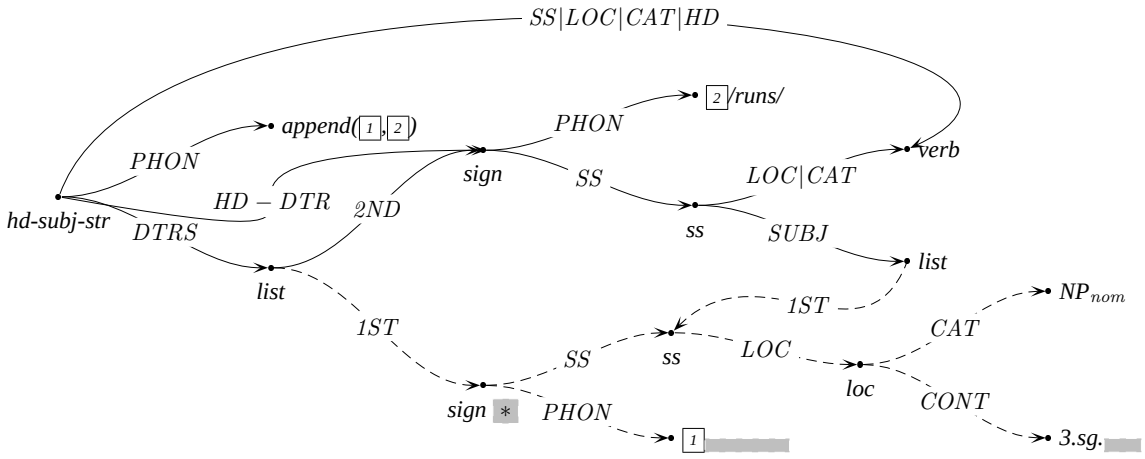


Figure 6: $TypeExp_T(F')$.

Suppose F is the feature structure in Figure 2 and D_F is its *n-lx* daughter. If D_F is removed from F it gives rise to a structure F' like the one in Figure 5. F' is type-expanded to $TypeExp$ (F') as in Figure 6.

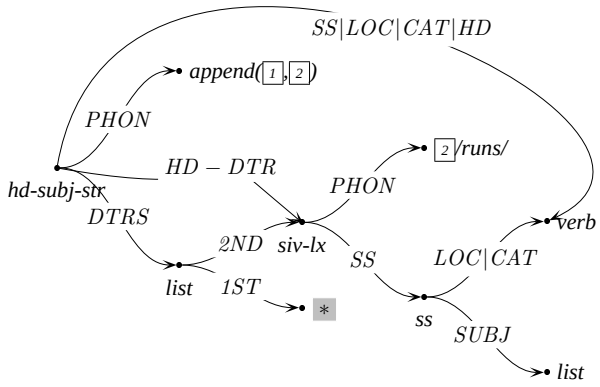


Figure 5: F': Erasing D_F.

On standard assumptions, any HPSG type theory for English will require that what fills the $DTRS|1ST$ slot be an object of type sign, whose $PHON$ is re-entrant with the first part of the $PHON$ of the whole sentence (i.e. tag 1) which follows from general constraints on *head-subj-str*). In addition, the *sign's* $SYNSEM$ *(SS)* value will be reentrant with the the $SUBJ|1ST$ slot of *runs*, which restricts it to being a 3rd person, singular, nominative nominal. There will be no constraint, however, that requires the subject of *runs* to be *fem*. $Context$ (D_F) denotes the subgraph rooted at *sign** in Figure 6 which is roughly $\text{NP}_{nom.3.sg}$.

Intuitively, these feature-value pairs could result from *Jane* being in that particular context (e.g. it might be that *3.sg* results from *Jane* being the subject of *runs*. $Inherent$ (D_F) denotes the relative complement of D_F and $Context$ (D_F) as depicted in Figure 7. It includes features that cannot reside in the context such as inherent phonological and semantic features of the entity in question, notably that its phonologocal content is */Jane/* and that it is feminine.

$$\begin{bmatrix} \textit{n-lx} \\ \text{PHON} \ \ /\text{Jane}/ \\ \text{SS} \mid \text{LOC} \mid \text{CONT} \quad \textit{fem} \end{bmatrix}$$

Figure 7: $Inherent$ (D_F)

Definition 2.2 ($Root$**)** *Given a representation F licensed by a type theory T, $Root$ selects any descendant D_F of F and returns $TypeExp_T(Inherent\ (D_F))$.*

Suppose, for example, $Root$ applies to *Jane* (i.e. the value of $DTRS|1ST$) in Figure 2. It will return $TypeExp_T(Inherent$ (*Jane*)), i.e. the type-expansion of the structure in Figure 7 (depicted in Figure 8) which has the properties of being nominal, and 3rd person singular but not nominative. The case restriction does not form part of $TypeExp_T(Inherent$ (*Jane*)) since there is nothing in the type theory to force it to 'grow back'. Notice that this fragment is of the right level of generality. Unlike the corresponding LFG-DOP fragment for *Jane*

in example (1)a, this is *3.sg.fem*, but not *nom*.[1]

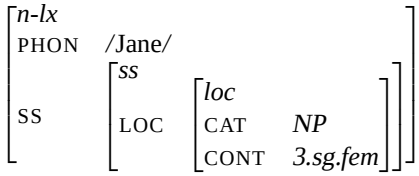


Figure 8: *TypeExp*$_T$(*Inherent* (*Jane*)).

Definition 2.3 (*Frontier***)** *Frontier erases any combination of F's descendant's and type expands the result F' marking the erasure points for composition.*[2]

If *Frontier* applies to *Jane* (i.e. the value of $DTRS|1ST$) in Figure 2, it will first erase the substructure corresponding to *Jane* as in Figure 5 marking the erasure point with $*$ and then type-expand the result (Figure 6). Notice *Frontier* produces again fragments of the right level of generality (i.e. general enough to allow both masculine and feminine subjects, but not sufficiently general to allow accusative subjects (e.g. **Him runs*) as was the case for LFG-DOP in example (2). Another fact about*Root* and *Frontier* as formulated here is that they do not require *Discard* to generalise over the fragments they produce, which is what causes the size of the fragment corpus to explode in the case of LFG-DOP.

2.3 Head-driven Composition

Standard composition approaches in DOP are rightwards or incrementally rightwards directed (Bod, 1995; Neumann, 2003)). In the context of HPSG, however, it is interesting to consider a "head-driven" approach to composition, whereby it is the head chain of the derivation initial fragment that identifies the order in which expansion nodes are to be considered as composition sites. Starting from the bottom, the open slot nodes of an *active fragment* are unified with other fragments so that each node along the path leading from the head lexical anchor to the root of the feature structure dominates a *passive* subconstituent before the next node along the path is considered. Composition is bidirectional with the direction being identified at each step (rather than in some predefined manner) by the head chain of the fragment rooted at the node being considered. Such a process is of course reminiscent of head-driven parsing strategies (Proudian and Pollard, 1985; van Noord, 1997, etc.).

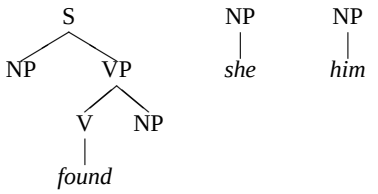


Figure 9 shows an example of head-driven composition in deriving a representation of *"She found him"* using the feature structure fragments corresponding to the subtrees above. The first internal node along the head chain of the derivation initial fragment (i.e. *hd-comp-struc*) dominates an active subconstituent. The rightmost terminal node is, therefore, the first node to be expanded. Unification proves successful and since *hd-comp-struc* now dominates a passive constituent the pointer is advanced one step along the path of the head chain to the root node *hd-subj-struc*. This again dominates an active subconstituent, thus identifying the next composition site. Fragment unification is again successful producing the last feature structure in Figure 9. This representation is *totally well-typed* and is, therefore, valid.

2.4 Fragment Probabilities

As in other DOP models, an HPSG-DOP representation will typically have many different derivations, and any string may have many different representations. Assuming composition steps are treated as independent events, the probabilities of a derivation $d =< f_1, ..., f_n >$ and a final representation R with m derivations d_j are defined as in (4) and (5) respectively. In order to compute fragment probabilities, we will use Tree-DOP's relative frequency estima-

[1] Had *Root* been applied to a node possesing the phonology *she* it would have produced something that *is nom*.

[2] This definition produces fragments analogous to those of Tree-DOP. Taking into account, however, that node labels in HPSG-DOP are subtypes of sign which do not convey any syntactic information, one might want to formulate *Frontier* so that it cannot apply to the overall lexical head restricting fragments to a minimum of one lexical anchor in order to maintain some syntactic information in the fragment (Linardaki, 2006).

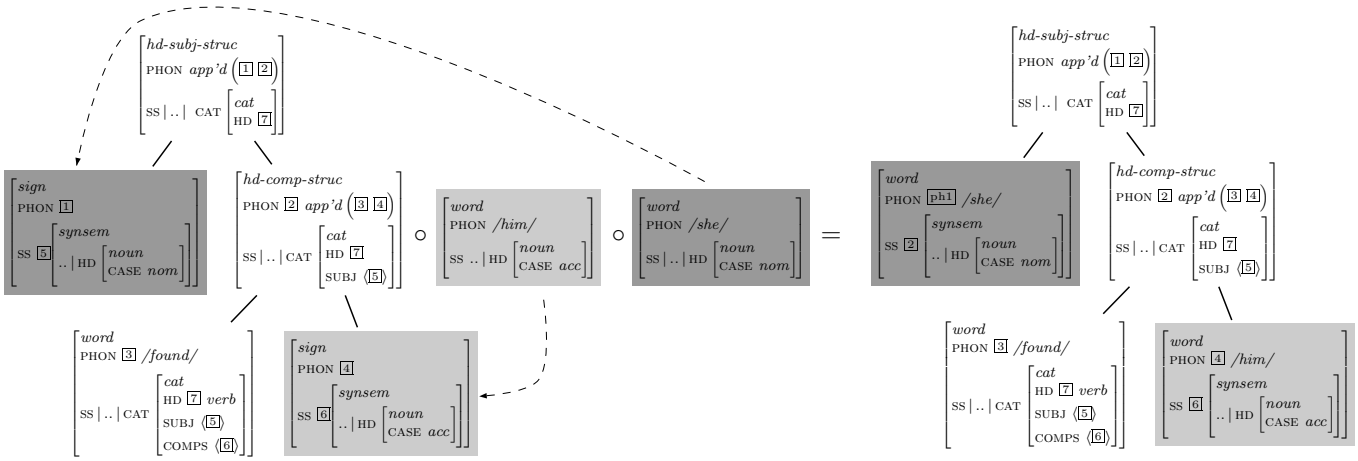


Figure 9: The head-driven composition operation.

tor as a starting point.

$$P(d) = \prod_{i=1}^{n} P(f_i) \tag{4}$$

$$P(R) = \sum_{j=1}^{m} P(d_j) \tag{5}$$

In Tree-DOP subtree probabilities are defined as in (6). The set of all composable subtrees at each derivation step is hence identified by its category $root(t_i)$. Category matching in HPSG-DOP corresponds to classifying fragments based on their head features and subcategorisation frame. Two fragments are hence considered as *competing* if they share the same values for all CAT features (i.e. HEAD, SUBJ, SPR and COMPS). The probability of a fragment f_i is then defined as in (7). Underspecified HEAD values are expanded in all possible ways and the resulting fragments are classified accordingly.

Presented at 'Exemplar-Based Models of Language Acquisition and Use'
Workshop, ESSLLI'07, Dublin, Aug 2007.

$$P(f_i) = \frac{|f_i|}{\sum_{root(f)=root(f_i)} |f|} \tag{6}$$

$$P(f_i) = \frac{|f_i|}{\sum_{\substack{v(\text{SS|LOC|CAT},r(f))= \\ v(\text{SS|LOC|CAT},r(f_i))}} |f|} \tag{7}$$

This stochastic process, however, is not guaranteed to identify a probability distribution over the set of valid representations because the combinatory potential of a fragment is not entirely determined by its CAT value. As a result, this process assigns some probability mass to structures outside the parse space (as is the case of category-identifiable sample spaces in LFG-DOP) and, hence, probability leak is observed.

Abney (1997) argues that relative frequency estimation constitutes a nonoptimal approach to probabilistic Attribute-Value Grammars (AVGs) in general due to the independence assumption not being applicable to deep linguistic analyses because their fragments are equipped to handle both syntactic and semantic dependencies. Loglinear or maximum entropy models (Abney, 1997; Miyao and Tsujii, 2002) are generally deemed as more suitable for such formalisms because they do not rely on the independence assumption.

The problem can be avoided by allowing competition sets to include all fragments that can be successfully unified with the next composition site (NCS) of some other fragment. Since previous derivation steps can affect the specificity of such sites, competition sets in HPSG-DOP cannot be predetermined. Suppose f_{i-1} is the structure produced before the i^{th} step of the derivation process. The probability of the next fragment to be used is defined as in (8). Fragment probabilities for the derivation initial selection can be based on category matching relative frequency estimation since there are no previ-

ous derivation steps to determine unifiability.

(8) $$P(f_i) = \frac{|f_i|}{\sum_{\substack{f \text{ is unifiable} \\ \text{with NCS}(f_{i-1})}} |f|}$$

3 Discussion

HPSG-DOP enjoys a number of positive characteristics. The most salient of these is its great linguistic sensitivity. It takes full advantage of the signature thus enabling the fragments produced to extend their ability of capturing dependencies beyond the syntactic level. HPSG-DOP's linguistic power, however, relies on grammaticality being defined entirely in terms of the signature. This is nearly, but not quite, true in standard HPSG. One example of information being determined outside the type theory is the nominal reference of *exempt* anaphors.

Principle A of the Binding theory states that a locally a-commanded anaphor must be locally a-bound. This implies that anaphors that are not locally a-commanded need not be a-bound. In (9) - (10), for example, *himself* and *themselves* are not a-commanded because the ARG-ST value of "picture" contains only one element (i.e. a PP). Such anaphors are known as exempt because they are exempted from the binding conditions.

(9) *John$_i$ took a picture of himself$_i$.*

(10) *They$_i$ saw pictures of themselves$_i$.*

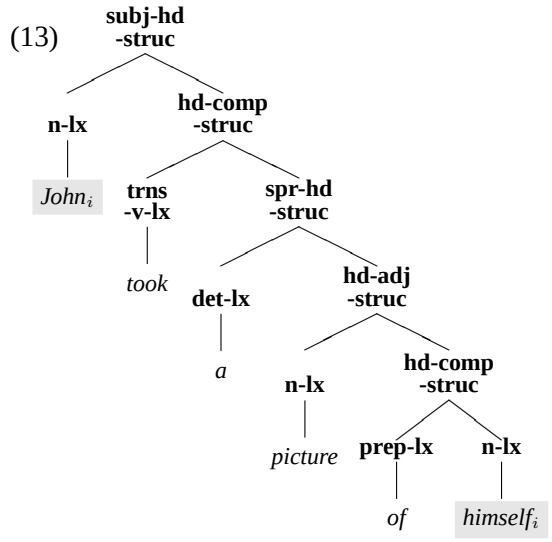
The implication of "need not" is what makes identifying the nominal reference of such anaphors go astray, because it does not determine whether *exempt anaphors* are, in fact, *a-bound* or not, and if yes to what. Consequently, the type theory in these cases licences more than what is intuitively correct. In the case of *"John took a picture of himself"*, for example, *"John"* does not have to be coindexed with *"himself"*, so (11) is perfectly acceptable for the type theory and, for the same reason, so is (12).

(11) **John$_i$ took a picture of himself$_j$.*

(12) **Mary took a picture of himself.*

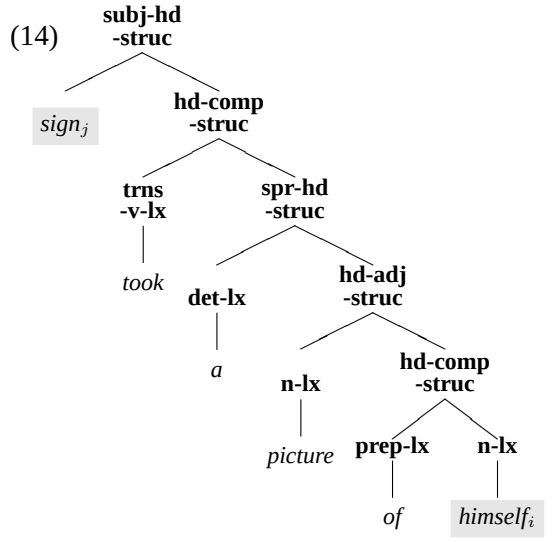
Next we will examine the effect of this in the context of HPSG-DOP. Take a simple training corpus containing the representation of the sentence *"John$_i$ took a picture of himself$_i$"* in (13).

(13)



Applying *Frontier* to the NP$_{John}$ node will produce the structure in (14). The implications of this are quite serious. Even though the produced fragment is valid, recombining it with the fragment that was cut off (i.e. NP$_{John}$) will not result in the initial structure because there is no constraint in the signature to reenforce the reentrance. In fact the structure in (13) cannot be reproduced in any way.

(14)



The source of the problem discussed here is that the type theory "overgenerates". One way of looking at this situation is hence to adopt the HPSG point

of view according to which the occurrence of *a-free* anaphors is linguistically valid so long as they are exempt. Even though this argument may be defensible from a linguistic point of view, it constitutes a less than satisfactory solution from the data-oriented point of view. The fact that no matter how much we overtrain a fragment corpus we will never be able to capture these dependencies when analysing new input stands in sharp contradiction with the DOP philosophy.

4 Concluding Remarks

We have presented a DOP model based on the syntactically and semantically more articulated representations of HPSG. The general architecture portrayed here allows for various HPSG-DOP instantiations, which will typically differ in the degree of specificity the fragments are allowed to have and/or the stochastic process employed for disambiguation.

Apart from the advantages that follow from the richer representational basis of such a model, it has a number of attractions. The most salient of these is its great linguistic sensitivity. It takes full advantage of the signature thus enabling the fragments produced to extend their ability of capturing dependencies beyond the syntactic level while at the same time they are of the right level of generality. Additionally, well-formedness of the final representation is checked during the derivation process. Even though, theoretically, this is not guarranteed, in practice the result of successful composition of totally well-typed feature structures will itself be a totally well-typed feature structure (i.e. a valid representation). As a result, relative frequency estimation can provide a feasible basis for computing the most probable analysis in HPSG-DOP, unlike other statistically enriched unification-based models.

HPSG-DOP's linguistic power, however, relies on grammaticality being defined entirely in terms of the signature. Unfortunately, this is not always the case in standard HPSG, where a number of phenomena are described outside the signature. In such cases decomposing an initial represenation and composing it again is not guarranteed to reproduce the same structure.

A natural objection from a language engineering point of view is that one of the attractions of the data-oriented philosophy (that it seems to dispense with the need to write grammars - all that is needed is a treebank) has been lost. Our approach lacks this attraction because it relies on the existence of a type system (i.e. an HPSG grammar). From a theoretical point of view, however, it is reasonable to have both a performance model (i.e. a form of DOP), and a competence grammar (i.e. a type theory).

Another positive data-oriented characteristic which has been sacrificed in order to ensure fragments are of the right level of generality is the feature of *robustness* (i.e. the ability to deal with input which is in some way ill-formed or extra-grammatical). This issue can be approached in a number of ways. In the case of unknown words, for example, the techniques described for Tree-DOP by (Bod, 1995) can be straightforwardly extended to this model. Robust unification (Fouvry, 2003), which is based on extending the signature to a lattice to include the unique joins of every set of incompatible types, provides a promising alternative to this issue. While we have discussed how HPSG-DOP behaves from a theoretical point of view, its empirical evaluation remains outstanding.

References

Steven P. Abney. 1997. Stochastic Attribute-Value Grammars. *Computational Linguistics*, 23(4):597–618.

Rens Bod and Ronald Kaplan. 1998. A Probabilistic Corpus-Driven Model for Lexical Functional Analysis. In *Proceedings of COLING-ACL'98*, Montreal, Canada.

Rens Bod. 1992. A Computational Model of Language Performance: Data Oriented Parsing. In *Proceedings COLING'92*, Nantes, France.

Rens Bod. 1995. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. Ph.D. thesis, Universiteit van Amsterdam, Amsterdam, The Netherlands.

Rens Bod. 2003. An Effi cient Implementation of a New DOP Model. In *Proceedings of the EACL 2003*, pages 19–26, Budapest, Hungary, April.

Frederik Fouvry. 2003. *Robust Processing for Constraint-based Grammar Formalisms*. Ph.D. thesis, University of Essex, April.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations*. CSLI Publications, Stanford.

Cornell Juliano and Michael K. Tanenhaus. 1993. Contingent Frequency Effects in Syntactic Ambiguity Resolution. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, University of Colorodo. Laurence Erlbaum Associates.

Daniel Jurafsky. 1996. A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science*, 20:137–194.

Evita Linardaki. 2006. *Linguistic and statistical extensions of Data Oriented Parsing*. Ph.D. thesis, University of Essex, UK.

Yusuke Miyao and Jun'ichi Tsujii. 2002. Maximum entropy estimation for feature forests. In *Proceedings of the Human Language Technology Conference*.

Günter Neumann. 1999. Learning Stochastic Lexicalized Tree Grammars from HPSG. Technical report, DFKI, Saarbruecken.

Günter Neumann. 2003. A Data-Driven Approach to Head-Driven Phrase Structure Grammar. In R. Scha R. Bod and K. Sima'an, editors, *Data-Oriented Parsing*, pages 233–251. CSLI Publications.

Carl J. Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago University Press.

Derek Proudian and Carl J. Pollard. 1985. Parsing head-driven phrase structure grammar. In *Proceedings of the Twenty-Third Annual Meeting of the ACL*, pages 167–171, Chicago, IL. ACL.

Gertjan van Noord. 1997. An efficient implementation of the head-corner parser. *Computational Linguistics*, 23(3):425–456.