# Designing Agreement Features for Realization Ranking

**Rajakrishnan Rajkumar** and **Michael White**
Department of Linguistics
The Ohio State University
{raja,mwhite}@ling.osu.edu

## Abstract

This paper shows that incorporating linguistically motivated features to ensure correct animacy and number agreement in an averaged perceptron ranking model for CCG realization helps improve a state-of-the-art baseline even further. Traditionally, these features have been modelled using hard constraints in the grammar. However, given the graded nature of grammaticality judgements in the case of animacy we argue a case for the use of a statistical model to rank competing preferences. Though subject-verb agreement is generally viewed to be syntactic in nature, a perusal of relevant examples discussed in the theoretical linguistics literature (Kathol, 1999; Pollard and Sag, 1994) points toward the heterogeneous nature of English agreement. Compared to writing grammar rules, our method is more robust and allows incorporating information from diverse sources in realization. We also show that the perceptron model can reduce balanced punctuation errors that would otherwise require a post-filter. The full model yields significant improvements in BLEU scores on Section 23 of the CCGbank and makes many fewer agreement errors.

## 1 Introduction

In recent years a variety of statistical models for realization ranking that take syntax into account have been proposed, including generative models (Bangalore and Rambow, 2000; Cahill and van Genabith, 2006; Hogan et al., 2007; Guo et al., 2008), maximum entropy models (Velldal and Oepen, 2005; Nakanishi et al., 2005) and averaged perceptron models (White and Rajkumar, 2009). To our knowledge, however, none of these models have included features specifically designed to handle grammatical agreement, an important task in surface realization. In this paper, we show that incorporating linguistically motivated features to ensure correct animacy and verbal agreement in an averaged perceptron ranking model for CCG realization helps improve a state-of-the-art baseline even further. We also demonstrate the utility of such an approach in ensuring the correct presentation of balanced punctuation marks.

Traditionally, grammatical agreement phenomena have been modelled using hard constraints in the grammar. Taking into consideration the range of acceptable variation in the case of animacy agreement and facts about the variety of factors contributing to number agreement, the question arises: tackle agreement through grammar engineering, or via a ranking model? In our experience, trying to add number and animacy agreement constraints to a grammar induced from the CCGbank (Hockenmaier and Steedman, 2007) turned out to be surprisingly difficult, as hard constraints often ended up breaking examples that were working without such constraints, due to exceptions, sub-regularities and acceptable variation in the data. With sufficient effort, it is conceivable that an approach incorporating hard agreement constraints could be refined to underspecify cases where variation is acceptable, but even so, one would want a ranking model to capture preferences in these cases, which might vary depending on genre, dialect or domain. Given that



*Coling 2010: Poster Volume*, pages 1032–1040,
Beijing, August 2010

a ranking model is desirable in any event, we investigate here the extent to which agreement phenomena can be more robustly and simply handled using a ranking model alone, with no hard constraints in the grammar.

We also show here that the perceptron model can reduce balanced punctuation errors that would otherwise require a post-filter. As White and Rajkumar (2008) discuss, in CCG it is not feasible to use features in the grammar to ensure that balanced punctuation (e.g. paired commas for NP appositives) is used in all and only the appropriate places, given the word-order flexibility that crossing composition allows. While a post-filter is a reasonably effective solution, it can be prone to search errors and does not allow balanced punctuation choices to interact with other choices made by the ranking model.

The starting point for our work is a CCG realization ranking model that incorporates Clark & Curran's (2007) normal-form syntactic model, developed for parsing, along with a variety of $n$-gram models. Although this syntactic model plays an important role in achieving top BLEU scores for a reversible, corpus-engineered grammar, an error analysis nevertheless revealed that many errors in relative pronoun animacy agreement and subject-verb number agreement remain with this model. In this paper, we show that features specifically designed to better handle these agreement phenomena can be incorporated into a realization ranking model that makes many fewer agreement errors, while also yielding significant improvements in BLEU scores on Section 23 of the CCGbank. These features make use of existing corpus annotations — specifically, PTB function tags and BBN named entity classes (Weischedel and Brunstein, 2005) — and thus they are relatively easy to implement.

### 1.1 The Graded Nature of Animacy Agreement

To illustrate the variation that can be found with animacy agreement phenomena, consider first animacy agreement with relative pronouns. In English, an inanimate noun can be modified by a relative clause introduced by *that* or *which*, while an animate noun combines with *who(m)*. With some nouns though — such as *team*, *group*, *squad*, etc. — animacy status is uncertain, and these can be found with all the three relative pronouns (*who, which* and *that*). Google counts suggest that all three choices are almost equally acceptable, as the examples below illustrate:

(1) The groups who protested against plans to remove asbestos from the nuclear submarine base at Faslane claimed victory when it was announced the government intends to dispose of the waste on site. (The Glasgow Herald; Jun 25, 2010)

(2) Mr. Dorsch says the HIAA is working on a proposal to establish a privately funded reinsurance mechanism to help cover small groups that ca n't get insurance without excluding certain employees . (WSJ0518.35)

### 1.2 The Heterogeneous Nature of Number Agreement

Subject-verb agreement can be described as a constraint where the verb agrees with the subject in terms of agreement features (number and person). Agreement has often been considered to be a syntactic phenomenon and grammar implementations generally use syntactic features to enforce agreement constraints (e.g. Velldal and Oepen, 2005). However a closer look at our data and a survey of the theoretical linguistics literature points toward a more heterogeneous conception of English agreement. Purely syntactic accounts are problematic when the following examples are considered:

(3) Five miles is a long distance to walk. (Kim, 2004)

(4) King prawns cooked in chili salt and pepper was very much better, a simple dish succulently executed. (Kim, 2004)

(5) " I think it will shake confidence one more time , and a lot of this business is based on client confidence . " (WSJ1866.10)

(6) It 's interesting to find that a lot of the expensive wines are n't always walking out the door . (WSJ0071.53)

In Example (3) above, the subject and determiner are plural while the verb is singular. In (4), the singular verb agrees with the dish, rather than with individual prawns. Measure nouns such as *lot*, *ton*, etc. exhibit singular agreement with the determiner *a*, but varying agreement with the verb depending on the head noun of the measure noun's *of*-complement. As is also well known, British and American English differ in subject-verb agreement with collective nouns. Kathol (1999) proposes an explanation where agreement is determined by the semantic properties of the noun rather than by its morphological properties. This accounts for all the cases above. In the light of this explanation, specifying agreement features in the logical form for realization could perhaps solve the problem. However, the semantic view of agreement is not completely convincing due to counterexamples like the following discussed in the literature (reported in Kim (2004)):

(7) Suppose you meet someone and they are totally full of themselves

(8) Those scissors are missing.

In Example (7), the pronoun *they* used in a generic sense is linked to the singular antecedent *someone*, but its plural feature triggers plural agreement with the verb. Example (8) illustrates a situation where the subject *scissors* is arguably semantically singular, but exhibits plural morphology and plural syntactic agreement with both the determiner as well as the verb. Thus this suggests that English has a set of heterogeneous agreement patterns rather than purely syntactic or semantic ones. This is also reflected in the proposal for a hybrid agreement system for English (Kim, 2004), where the morphology tightly interacts with the system of syntax, semantics, or even pragmatics to account for agreement phenomena. Our machine learning-based approach approximates the insights discussed in the theoretical linguistics literature. Writing grammar rules to get these facts right proved to be surprisingly difficult (e.g. discerning the actual nominal head contributing agreement feature in cases like *areas of the factory were/*was* vs. *a lot of wines are/*is*) and required a list of measure nouns and partitive quantifiers. We investigate here the extent to which a machine learning–based approach is a simpler, practical alternative for acquiring the relevant generalizations from the data by combining information from various information sources.

The paper is structured as follows. Section 2 provides CCG background. Section 3 describes the features we have designed for animacy and number agreement as well as for balanced punctuation. Section 4 presents our evaluation of the impact of these features in averaged perceptron realization ranking models, tabulating specific kinds of errors in the CCGbank development section as well as overall automatic metric scores on Section 23. Section 5 compares our results to those obtained with related systems. Finally, Section 6 concludes with a summary of the paper's contributions.

## 2 Background

### 2.1 Surface Realization with Combinatory Categorial Grammar (CCG)

CCG (Steedman, 2000) is a unification-based categorial grammar formalism which is defined almost entirely in terms of lexical entries that encode sub-categorization information as well as syntactic feature information (e.g. number and agreement). Complementing function application as the standard means of combining a head with its argument, type-raising and composition support transparent analyses for a wide range of phenomena, including right-node raising and long distance dependencies. An example syntactic derivation appears in Figure 1, with a long-distance dependency between *point* and *make*. Semantic composition happens in parallel with syntactic composition, which makes it attractive for generation.

OpenCCG is a parsing/generation library which works by combining lexical categories for words using CCG rules and multi-modal extensions on rules (Baldridge, 2002) to produce derivations. Conceptually these extensions are on lexical categories. Surface realization is the process by which logical forms are transduced to strings. OpenCCG uses a hybrid symbolic-statistical chart realizer (White, 2006) which takes logical forms as input and produces sentences by using CCG com-

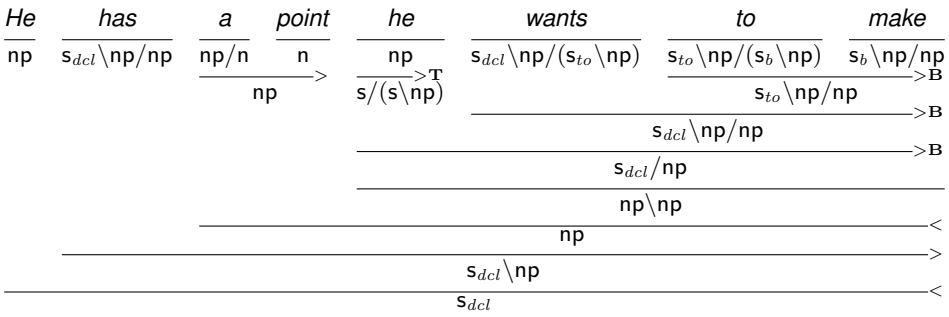


Figure 1: Syntactic derivation from the CCGbank for *He has a point he wants to make [. . . ]*

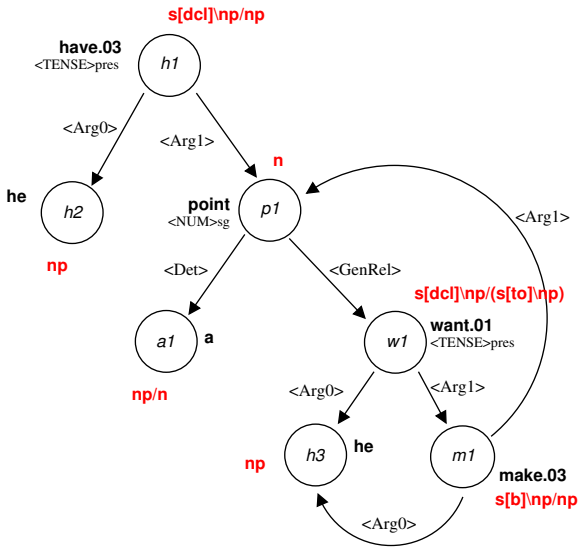


Figure 2: Semantic dependency graph from the CCGbank for *He has a point he wants to make [. . . ]*, along with gold-standard supertags (category labels)

binators to combine signs. Edges are grouped into equivalence classes when they have the same syntactic category and cover the same parts of the input logical form. Alternative realizations are ranked using integrated $n$-gram or perceptron scoring, and pruning takes place within equivalence classes of edges. To more robustly support broad coverage surface realization, OpenCCG greedily assembles fragments in the event that the realizer fails to find a complete realization.

To illustrate the input to OpenCCG, consider the semantic dependency graph in Figure 2. In the graph, each node has a lexical predication (e.g. **make.03**) and a set of semantic features (e.g. ⟨NUM⟩sg); nodes are connected via dependency relations (e.g. ⟨ARG0⟩). (Gold-standard supertags, or category labels, are also shown; see Section 2.2 for their role in hypertagging.) Internally, such graphs are represented using Hybrid Logic Dependency Semantics (HLDS), a dependency-based approach to representing linguistic meaning (Baldridge and Kruijff, 2002). In HLDS, each semantic head (corresponding to a node in the graph) is associated with a nominal that identifies its discourse referent, and relations between heads and their dependents are modeled as modal relations.

For our experiments, we use an enhanced version of the CCGbank (Hockenmaier and Steedman, 2007)—a corpus of CCG derivations derived from the Penn Treebank—with Propbank (Palmer et al., 2005) roles projected onto it (Boxwell and White, 2008). Additionally, certain multi-word NEs were collapsed using underscores so that they are treated as atomic entities in the input to the realizer. To engineer a grammar from this corpus suitable for realization with OpenCCG, the derivations are first revised to reflect the lexicalized treatment of coordination and punctuation assumed by the multi-modal version of CCG that is implemented in OpenCCG (White and Rajkumar, 2008). Further changes are necessary to support semantic dependencies rather than surface syntactic ones; in particular, the features and unification constraints in the categories related to semantically empty function words such complementizers, infinitival-*to*, expletive subjects, and case-marking prepositions are adjusted to reflect their purely syntactic status.

### 2.2 Hypertagging

A crucial component of the OpenCCG realizer is the *hypertagger* (Espinosa et al., 2008), or supertagger for surface realization, which uses a maximum entropy model to assign the most likely lexical categories to the predicates in the input logical form, thereby greatly constraining the realizer's search space.[1] Category label prediction is done at run-time and is based on contexts within the directed graph structure as shown in Figure 2, instead of basing category assignment on linear word and POS context as in the parsing case.

## 3 Feature Design

The features we employ in our baseline perceptron ranking model are of three kinds. First, as in the log-linear models of Velldal & Oepen and Nakanishi et al., we incorporate the log probability of the candidate realization's word sequence according to our linearly interpolated language models as a single feature in the perceptron model. Since our language model linearly interpolates three component models, we also include the log prob from each component language model as a feature so that the combination of these components can be optimized. Second, we include syntactic features in our model by implementing Clark & Curran's (2007) normal form model in OpenCCG. The features of this model are listed in Table 1; they are integer-valued, representing counts of occurrences in a derivation. Third, we include discriminative $n$-gram features (Roark et al., 2004), which count the occurrences of each $n$-gram that is scored by our factored language model, rather than a feature whose value is the log probability determined by the language model. Table 2 depicts the new animacy, agreement and punctuation features being introduced as part of this work. The next two sections describe these features in more detail.

### 3.1 Animacy and Number Agreement

Underspecification as to the choice of pronoun in the input leads to competing realizations involving the relative pronouns *who*, *that*, *which* etc. The existing ranking models ($n$-gram models as well as perceptron) often allow the top-ranked output to have the relative pronoun *that* associated with animate nouns. The existing normal form model uses the word forms as well as part-of-speech tag based features. Though this is useful for associating proper nouns (tagged NNP or NNPS) with *who*, for other nouns (as in *consumers who* vs. *consumers that/which*), the model often prefers the infelicitous pronoun. So here we designed features which also took into account the named entity class of the head noun as well as the stem of the head noun. These features aid the discriminative $n$-gram features (*PERSON, which* has high negative weight). As the results section discusses,

[1] The approach has been dubbed *hypertagging* since it operates at a level "above" the syntax, moving from semantic representations to syntactic categories.

| Feature Type | Example |
|---|---|
| LexCat + Word | $s/s/np$ + before |
| LexCat + POS | $s/s/np$ + IN |
| Rule | $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$ |
| Rule + Word | $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$ + bought |
| Rule + POS | $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$ + VBD |
| Word-Word | $\langle$company, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, bought$\rangle$ |
| Word-POS | $\langle$company, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, VBD$\rangle$ |
| POS-Word | $\langle$NN, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np$, bought$\rangle$ |
| Word + $\Delta_w$ | $\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_w$ |
| POS + $\Delta_w$ | $\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_w$ |
| Word + $\Delta_p$ | $\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_p$ |
| POS + $\Delta_p$ | $\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_p$ |
| Word + $\Delta_v$ | $\langle$bought, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_v$ |
| POS + $\Delta_v$ | $\langle$VBD, $s_{dcl} \rightarrow np\ s_{dcl}\backslash np\rangle + d_v$ |

Table 1: Baseline features: Basic and dependency features from Clark & Curran's (2007) normal form model; distances are in intervening words, punctuation marks and verbs, and are capped at 3, 3 and 2, respectively

| Feature | Example |
|---|---|
| **Animacy features** | |
| Noun Stem + Wh-pronoun | researcher + who |
| Noun Class + Wh-pronoun | PER_DESC + who |
| **Number features** | |
| Noun + Verb | people + are |
| NounPOS + Verb | NNS + are |
| Noun + VerbPOS | people + VBP |
| NounPOS + VerbPOS | NNS + VBP |
| Noun_of + Verb | lot_of + are |
| Noun_of + VerbPOS | lot_of + VBP |
| NounPOS_of + Verb | NN_of + are |
| NounPOS_of + VerbPOS | NN_of + VBP |
| Noun_of + *of*-complementPOS + VerbPOS | lot_of + NN + VBZ |
| NounPOS_of + *of*-complementPOS + VerbPOS | NN_of + NN + VBZ |
| Noun_of + *of*-complementPOS + Verb | lot_of + NN + is |
| NounPOS_of + *of*-complementPOS + Verb | NN_of + NN + is |
| **Punctuation feature** | |
| Balanced Punctuation Indicator | $unbalPunct=1 |

Table 2: New features introduced

NE classes like PER_DESC contribute substantially towards animacy preferences.

For number agreement, we designed three classes of features (c.f. *Number Agr* row in Table 2). Each of these classes results in 4 features. During feature extraction, subjects of the verbs tagged VBZ and VBP and verbs *was, were* were identified using the PTB NP-SBJ function tag annotation projected on to the appropriate arguments of lexical categories of verbs. The first class of features encoded all possible combinations of subject-verb word forms and parts of speech tags. In the case of NPs involving *of*-complements like *a lot of ...* (Examples 5 and 6), feature classes 2 and 3 were extracted (class 1 was excluded). Class 2 features encode the fact that the syntactic head has an associated *of*-complement, while class 3 features also include the part of speech tag of the complement. In the case of conjunct/disjunct VPs and subject NPs, the feature specifically looked at the parts of speech of both the NPs/VPs forming the conjunct/disjunct. The motivation behind such a design was to glean syntactic and semantic generalizations from the data. During feature extraction, from each derivation, counts of animacy and agreement features were obtained.

### 3.2 Balanced Punctuation

A complex issue that arises in the design of bidirectional grammars is ensuring the proper presentation of punctuation. Among other things, this involves the task of ensuring the correct realization of commas introducing noun phrase appositives.

(9) John, CEO of ABC, loves Mary.
(10) * John, CEO of ABC loves Mary.
(11) Mary loves John, CEO of ABC.
(12) * Mary loves John, CEO of ABC,.
(13) Mary loves John, CEO of ABC, madly.
(14) * Mary loves John, CEO of ABC madly.

As of now, $n$-gram models rule out examples like 12 above. All the other unacceptable examples are ruled out using a post-filter on realized derivations. As described in White and Rajkumar (2008), the need for the filter arises because a feature-based approach appears to be inadequate for dealing with the class of examples presented above in CCG. This approach involves the incorporation of syntactic features for punctuation into atomic categories so that certain combinations are blocked. To ensure proper appositive balancing sentence finally, the rightmost element in the sentence should transmit a relevant feature to the clause level, which the sentence-final period can then check for the presence of right-edge punctuation. However, the feature schema does not constrain cases of balanced punctuation in cases involving crossing composition and extraction. However, in this paper we explore a statistical approach to ensure proper balancing of NP apposition commas. The first step in this solution is the introduction of a feature in the grammar which indicates balanced vs. unbalanced marks. We modified the result categories of unbalanced appositive commas and dashes to include a feature marking unbalanced punctuation, as follows:

(15) $, \vdash \mathsf{np}_{\langle 1 \rangle unbal=comma} \backslash_\star \mathsf{np}_{\langle 1 \rangle} /_\star \mathsf{np}_{\langle 2 \rangle}$

Then, during feature extraction, derivations were examined to detect categories such as $\mathsf{np}_{unbal=comma}$, and checked to make sure this NP is followed by another punctuation mark in the string such as a full stop. The feature indicates the presence or absence of unbalanced punctuation in the derivation.

## 4 Evaluation

### 4.1 Experimental Conditions

For the experiments reported below, we used a lexico-grammar extracted from Sections 02–21 of our enhanced CCGbank with collapsed NEs, a hypertagging model incorporating named entity class features, and a trigram factored language model over words, named entity classes, part-of-speech tags and supertags. Perceptron training events were generated for each training section separately. The hypertagger and POS/supertag language model were trained on all the training sections, while separate word-based models were trained excluding each of the training sections in turn. Event files for 26530 training sentences with complete realizations were generated, with an average n-best list size of 18.2. The complete set of models is listed in Table 3.

| Model | Description |
|---|---|
| full-model | All the feats from models below |
| agr-punct | Baseline Feats + Punct + Num-Agr |
| wh-punct | Baseline Feats + Punct + Animacy-Agr |
| baseline-punct | Baseline Feats + Punct |
| baseline | Log prob + $n$-gram +Syntactic features |

Table 3: Legend for experimental conditions

### 4.2 Results

Realization results on the development and test sections are given in Table 4. For the development section, in terms of both exact matches and BLEU scores, the model with all the three features discussed above (agreement, animacy and punctuation) performs better than the baseline which does not have any of these features. However, using these criteria, the best performing model is actually the model which has agreement and punctuation features. The model containing all the features does better than the punctuation-feature only model, but performs slightly worse than the agreement-punctuation model. Section 23, the test section, confirms that the model with all the features performs better than the baseline model. We calculated statistical significance for the main results using bootstrap random sampling.[2] After re-sampling 1000 times, significance was calculated using a paired t-test (999 d.f.). The results indicated that the model with all the features in it (full-model) exceeded the baseline with $p < 0.0001$ . However, exact matches and BLEU scores do not necessarily reflect the extent to which important grammatical flaws have been reduced. So to judge the effectiveness of the new features, we computed the percentage of errors of each type that were present in the best Section 00 realization selected by each of these models. Also note that our baseline results differ slightly from the corresponding results reported in White and Rajkumar (2009) in spite of using the same feature set because quotes were introduced into the corpus on which these experiments were conducted. Previous results were based on the original CCGbank text where quotation marks are absent.

Table 6 reports results of the error analysis. It

[2]Scripts for running these tests are available at `http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm`

| Section | Model | %Exact | %Compl. | BLEU |
|---|---|---|---|---|
| 00 | baseline | 38.18 | 82.47 | 0.8341 |
| | baseline-punct | 37.97 | 82.47 | 0.8340 |
| | wh-punct | 38.93 | 82.53 | 0.8360 |
| | full-model | 40.47 | 82.53 | 0.8403 |
| | agr-punct | 40.84 | 82.53 | **0.8414** |
| 23 | baseline | 38.98 | 83.39 | 0.8442 |
| | full-model | 40.09 | 83.35 | **0.8446** |

Table 4: Results (98.9% coverage)—percentage of exact match and grammatically complete realizations and BLEU scores

| Model | METEOR | TERP |
|---|---|---|
| baseline | 0.9819 | 0.0939 |
| baseline-punct | 0.9819 | 0.0939 |
| wh-punct | 0.9827 | 0.0923 |
| agr-punct | 0.9821 | 0.0902 |
| full-model | 0.9826 | 0.0909 |

Table 5: Section 00 METEOR and TERP scores

can be seen that the punctuation-feature is effective in reducing the number of sentences with unbalanced punctuation marks. Similarly, the full model has fewer animacy mismatches and just about the same number of errors of the other two types, though it performs slightly worse than the agreement-only model in terms of BLEU scores and exact matches. We also manually examined the remaining cases of animacy agreement errors in the output of the full model here. Of the remaining 18 errors, 14 were acceptable paraphrases involving object relative clauses (eg. wsj_0083.40 ... *the business that*/∅ *a company can generate*). We also provide METEOR and TERP scores for these models (Table 5). In recently completed work on the creation of a human-rated paraphrase corpus to evaluate NLG systems, our analyses showed that BLEU, METEOR and TERP scores correlate moderately with human judgments of adequacy and fluency, and that the most reliable system-level comparisons can be made only by looking at all three metrics.

### 4.3 Examples

Table 7 presents four examples where the full model differs from the baseline. Example wsj_0003.8 illustrates an example where the NE tag PER_DESC for *researchers* helps the perceptron model enforce the correct animacy agreement, while the two baseline models prefer the

| | |
|---|---|
| Ref-wsj_0003.8 full,agr,wh | neither Lorillard nor the **researchers who** studied the workers were aware of any research on smokers of the Kent cigarettes |
| baseline,baseline-punct | neither Lorillard nor the **researchers that** studied the workers were aware of any research on smokers of the Kent cigarettes . |
| Ref-wsj_0003.18 agr-punct, full | the **plant** , which is owned by Hollingsworth & Vose Co. , **was** under contract with lorillard to make the cigarette filters . |
| baselines, wh | the **plant** , which is owned by Hollingsworth & Vose Co. , **were** under contract with lorillard to make the cigarette filters . |
| Ref-wsj_0018.6 agr-punct, full model | while **many of the risks were** anticipated when minneapolis-based Cray Research first announced the spinoff ... |
| agr-punct, full | while **many of the risks were** anticipated when minneapolis-based Cray Research first announced the spinoff ... |
| baselines | while **many of the risks was** anticipated when minneapolis-based Cray Research announced the spinoff ... |
| Ref-wsj_0070.4 agr-punct, full | Giant Group is led by three Rally 's directors , Burt Sugarman , James M. Trotter III and William E. Trotter II that last month indicated that **they hold** a 42.5 % stake in Rally 's and **plan** to seek a majority of seats on ... |
| all others | Giant Group is led by three Rally 's directors , Burt Sugarman , James M. Trotter III and William E. Trotter II that last month indicated that **they holds** a 42.5 % stake in Rally 's and **plans** to seek a majority of seats on ... |
| Ref-wsj_0047.5 | ... the ban wo n't stop privately funded tissue-transplant research or federally funded fetal-tissue research that does n't involve transplants . |
| agr, full | ... the ban wo n't stop tissue-transplant privately funded **research** or federally funded fetal-tissue **research** that **does n't** involve transplants . |
| baselines, wh | ... the ban wo n't stop tissue-transplant privately funded **research** or federally funded fetal-tissue **research** that **do n't** involve transplants . |

Table 7: Examples of realized output

| Model | #Punct-Errs | %Agr-Errs | %WH-Errs |
|---|---|---|---|
| baseline | 39 | 11.05 | 22.44 |
| baseline-punct | 0 | 10.79 | 20.77 |
| wh-punct | 11 | 10.87 | 13.53 |
| agr-punct | 8 | 4.0 | 21.84 |
| full-model | 10 | 4.31 | 15.53 |

Table 6: Error analysis of Section 00 complete realizations (total of 1554 agreement cases; total of 207 WH-pronoun cases)

*that* realization. Example wsj_0003.18 illustrates an instance of simple subject-verb agreement being enforced by the models containing the agreement features. Example wsj_0070.4 presents a more complex situation where a single subject has to agree with both verbs in a conjoined verb phrase. The last example in Table 7 shows the case of a NP subject which is a disjunction of two individual NPs. In both these cases, while the baseline models do not enforce the correct choice, the models with the agreement features do get this right. This is because our agreement features are sensitive to the properties of both NP and VP conjuncts/disjuncts. In addition, most of the realizations involving *of*-complements are also ranked correctly. In the final example sentence provided (i.e. wsj_0018.6), the models with the agreement features are able to enforce the correct the agreement constraints in the phrase *many of the risks were* in contrast to the baseline models.

## 5 Conclusion

In this paper, we have shown for the first time that incorporating linguistically motivated features to ensure correct animacy and number agreement in a statistical realization ranking model yields significant improvements over a state-of-the-art baseline. While agreement has traditionally been modelled using hard constraints in the grammar, we have argued that using a statistical ranking model is a simpler and more robust approach that is capable of learning competing preferences and cases of acceptable variation. Our approach also approximates insights about agreement which have been discussed in the theoretical linguistics literature. We have also shown how a targeted error analysis can reveal substantial reductions in agreement errors, whose impact on quality no doubt exceeds what is suggested by the small BLEU score increases. As future work, we also plan to learn such patterns from large amounts of unlabelled data and use models learned thus to rank paraphrases.

## Acknowledgements

This work was supported in part by NSF grant IIS-0812297 and by an allocation of computing time from the Ohio Supercomputer Center. Our thanks also to Robert Levine and the anonymous reviewers for helpful comments and discussion.

## References


Baldridge, Jason and Geert-Jan Kruijff. 2002. Coupling CCG and Hybrid Logic Dependency Semantics. In *Proc. ACL-02*.

Baldridge, Jason. 2002. *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Bangalore, Srinivas and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proc. COLING-00*.

Boxwell, Stephen and Michael White. 2008. Projecting Propbank roles onto the CCGbank. In *Proc. LREC-08*.

Cahill, Aoife and Josef van Genabith. 2006. Robust PCFG-based generation using automatically acquired LFG approximations. In *Proc. COLING-ACL '06*.

Clark, Stephen and James R. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33(4):493–552.

Espinosa, Dominic, Michael White, and Dennis Mehay. 2008. Hypertagging: Supertagging for surface realization with CCG. In *Proc. ACL-08: HLT*.

Guo, Yuqing, Josef van Genabith, and Haifeng Wang. 2008. Dependency-based n-gram models for general purpose sentence realisation. In *Proc. COLING-08*.

Hockenmaier, Julia and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Hogan, Deirdre, Conor Cafferkey, Aoife Cahill, and Josef van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *Proc. EMNLP-CoNLL*.

Kathol, Andreas. 1999. Agreement and the Syntax-Morphology Interface in HPSG. In Levine, Robert D. and Georgia M. Green, editors, *Studies in Contemporary Phrase Structure Grammar*, pages 223–274. Cambridge University Press, Cambridge.

Kim, Jong-Bok. 2004. Hybrid Agreement in English. *Linguistics*, 42(6):1105–1128.

Nakanishi, Hiroko, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic methods for disambiguation of an HPSG-based chart generator. In *Proc. IWPT-05*.

Palmer, Martha, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

Pollard, Carl and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. University Of Chicago Press.

Roark, Brian, Murat Saraclar, Michael Collins, and Mark Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proc. ACL-04*.

Steedman, Mark. 2000. *The syntactic process*. MIT Press, Cambridge, MA, USA.

Velldal, Erik and Stephan Oepen. 2005. Maximum entropy models for realization ranking. In *Proc. MT Summit X*.

Weischedel, Ralph and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. Technical report, BBN.

White, Michael and Rajakrishnan Rajkumar. 2008. A more precise analysis of punctuation for broad-coverage surface realization with CCG. In *Proc. of the Workshop on Grammar Engineering Across Frameworks (GEAF08)*.

White, Michael and Rajakrishnan Rajkumar. 2009. Perceptron reranking for CCG realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Singapore, August. Association for Computational Linguistics.

White, Michael. 2006. Efficient Realization of Coordinate Structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.