Underspecifying and Predicting Voice for Surface Realisation Ranking

Sina Zarrieß, Aoife Cahill and Jonas Kuhn
Institut für maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{sina.zarriess,aoife.cahill,jonas.kuhn}@ims.uni-stuttgart.de

Abstract

This paper addresses a data-driven surface realisation model based on a large-scale reversible grammar of German. We investigate the relationship between the surface realisation performance and the character of the input to generation, i.e. its degree of underspecification. We extend a syntactic surface realisation system, which can be trained to choose among word order variants, such that the candidate set includes active and passive variants. This allows us to study the interaction of voice and word order alternations in realistic German corpus data. We show that with an appropriately underspecified input, a linguistically informed realisation model trained to regenerate strings from the underlying semantic representation achieves 91.5% accuracy (over a baseline of 82.5%) in the prediction of the original voice.

1 Introduction

This paper[1] presents work on modelling the usage of voice and word order alternations in a free word order language. Given a set of meaning-equivalent candidate sentences, such as in the simplified English Example (1), our model makes predictions about which candidate sentence is most appropriate or natural given the context.

(1) Context: *The Parliament started the debate about the state budget in April.*

a. It wasn't until June that the Parliament approved it.
b. It wasn't until June that it was approved by the Parliament.
c. It wasn't until June that it was approved.

We address the problem of predicting the usage of linguistic alternations in the framework of a *surface realisation ranking* system. Such ranking systems are practically relevant for the real-world application of grammar-based generators that usually generate several grammatical surface sentences from a given abstract input, e.g. (Velldal and Oepen, 2006). Moreover, this framework allows for detailed experimental studies of the interaction of specific linguistic features. Thus it has been demonstrated that for free word order languages like German, word order prediction quality can be improved with carefully designed, linguistically informed models capturing information-structural strategies (Filippova and Strube, 2007; Cahill and Riester, 2009).

This paper is situated in the same framework, using rich linguistic representations over corpus data for machine learning of realisation ranking. However, we go beyond the task of finding the correct ordering for an almost fixed set of word forms. Quite obviously, word order is only one of the means at a speaker's disposal for expressing some content in a contextually appropriate form; we add systematic alternations like the voice alternation (active vs. passive) to the picture. As an alternative way of promoting or demoting the prominence of a syntactic argument, its interaction with word ordering strategies in real corpus data is of high theoretical interest (Aissen, 1999; Aissen, 2003; Bresnan et al., 2001).

Our main goals are (i) to establish a corpus-based surface realisation framework for empirically investigating interactions of voice and word order in German, (ii) to design an input representation for generation capturing voice alternations in a variety of contexts, (iii) to better understand the relationship between the performance of a generation ranking model and the type of realisation candidates available in its input. In working towards these goals, this paper addresses the question of evaluation. We conduct a pilot human evaluation on the voice al-

[1]This work has been supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in SFB 732 *Incremental specification in context*, project D2 (PIs: Jonas Kuhn and Christian Rohrer).

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1007–1017,
Portland, Oregon, June 19-24, 2011.

ternation data and relate our findings to our results established in the automatic ranking experiments.

Addressing interactions among a range of grammatical and discourse phenomena on realistic corpus data turns out to be a major methodological challenge for data-driven surface realisation. The set of candidate realisations available for ranking will influence the findings, and here, existing surface realisers vary considerably. Belz et al. (2010) point out the differences across approaches in the type of syntactic and semantic information present and absent in the input representation; and it is the type of underspecification that determines the number (and character) of available candidate realisations and, hence, the complexity of the realisation task.

We study the effect of varying degrees of underspecification explicitly, extending a syntactic generation system by a semantic component capturing voice alternations. In regeneration studies involving underspecified underlying representations, corpus-oriented work reveals an additional methodological challenge. When using standard semantic representations, as common in broad-coverage work in semantic parsing (i.e., from the point of view of analysis), alternative variants for sentence realisation will often receive slightly different representations: In the context of (1), the continuation (1-c) is presumably more natural than (1-b), but with a standard sentence-bounded semantic analysis, only (1-a) and (1-b) would receive equivalent representations.

Rather than waiting for the availability of robust and reliable techniques for detecting the reference of implicit arguments in analysis (or for contextually aware reasoning components), we adopt a relatively simple heuristic approach (see Section 3.1) that approximates the desired equivalences by augmented representations for examples like (1-c). This way we can overcome an extremely skewed distribution in the naturally occurring meaning-equivalent active vs. passive sentences, a factor which we believe justifies taking the risk of occasional overgeneration.

The paper is structured as follows: Section 2 situates our methodology with respect to other work on surface realisation and briefly summarises the relevant theoretical linguistic background. In Section 3, we present our generation architecture and the design of the input representation. Section 4 describes the setup for the experiments in Section 5. In Section 6, we present the results from the human evaluation.

2 Related Work

2.1 Generation Background

The first widely known data-driven approach to surface realisation, or tactical generation, (Langkilde and Knight, 1998) used language-model n-gram statistics on a word lattice of candidate realisations to guide a ranker. Subsequent work explored ways of exploiting linguistically annotated data for trainable generation models (Ratnaparkhi, 2000; Marciniak and Strube, 2005; Belz, 2005, a.o.). Work on data-driven approaches has led to insights into the importance of linguistic features for sentence linearisation decisions (Ringger et al., 2004; Filippova and Strube, 2009). The availability of discriminative learning techniques for the ranking of candidate analyses output by broad-coverage grammars with rich linguistic representations, originally in parsing (Riezler et al., 2000; Riezler et al., 2002), has also led to a revival of interest in linguistically sophisticated reversible grammars as the basis for surface realisation (Velldal and Oepen, 2006; Cahill et al., 2007). The grammar generates candidate analyses for an underlying representation and the ranker's task is to predict the contextually appropriate realisation.

The work that is most closely related to ours is Velldal (2008). He uses an MRS representation derived by an HPSG grammar that can be underspecified for information status. In his case, the underspecification is encoded in the grammar and not directly controlled. In multilingually oriented linearisation work, Bohnet et al. (2010) generate from semantic corpus annotations included in the CoNLL'09 shared task data. However, they note that these annotations are not suitable for full generation since they are often incomplete. Thus, it is not clear to which degree these annotations are actually underspecified for certain paraphrases.

2.2 Linguistic Background

In competition-based linguistic theories (Optimality Theory and related frameworks), the use of argument alternations is construed as an effect of markedness hierarchies (Aissen, 1999; Aissen, 2003). Argument functions (subject, object, ...) on

the one hand and the various properties that argument phrases can bear (person, animacy, definiteness) on the other are organised in markedness hierarchies. Wherever possible, there is a tendency to *align* the hierarchies, i.e., use prominent functions to realise prominently marked argument phrases. For instance, Bresnan et al. (2001) find that there is a statistical tendency in English to passivise a verb if the patient is higher on the person scale than the agent, but an active is grammatically possible.

Bresnan et al. (2007) correlate the use of the English dative alternation to a number of features such as givenness, pronominalisation, definiteness, constituent length, animacy of the involved verb arguments. These features are assumed to reflect the discourse acessibility of the arguments.

Interestingly, the properties that have been used to model argument alternations in strict word order languages like English have been identified as factors that influence word order in free word order languages like German, see Filippova and Strube (2007) for a number of pointers. Cahill and Riester (2009) implement a model for German word order variation that approximates the information status of constituents through morphological features like definiteness, pronominalisation etc. We are not aware of any corpus-based generation studies investigating how these properties relate to argument alternations in free word order languages.

3 Generation Architecture

Our data-driven methodology for investigating factors relevant to surface realisation uses a regeneration set-up[2] with two main components: a) a grammar-based component used to parse a corpus sentence and map it to all its meaning-equivalent surface realisations, b) a statistical ranking component used to select the correct, i.e. contextually most appropriate surface realisation. Two variants of this set-up that we use are sketched in Figure 1.

We generally use a hand-crafted, broad-coverage LFG for German (Rohrer and Forst, 2006) to parse a corpus sentence into a f(unctional) structure[3] and generate all surface realisations from a given

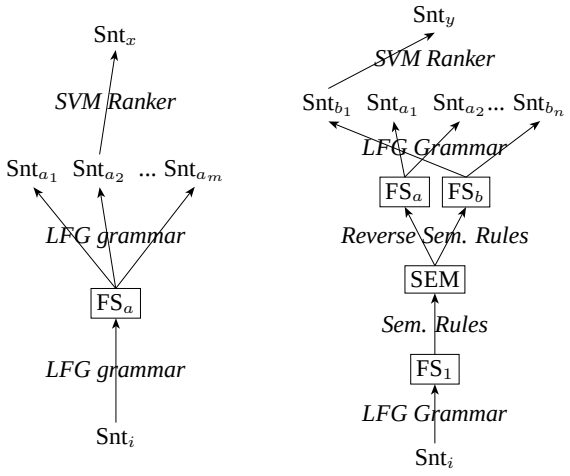

Figure 1: Generation pipelines

f-structure, following the generation approach of Cahill et al. (2007). F-structures are attribute-value matrices representing grammatical functions and morphosyntactic features; their theoretical motivation lies in the abstraction over details of surface realisation. The grammar is implemented in the XLE framework (Crouch et al., 2006), which allows for reversible use of the same declarative grammar in the parsing and generation direction.

To obtain a more abstract underlying representation (in the pipeline on the right-hand side of Figure 1), the present work uses an additional semantic construction component (Crouch and King, 2006; Zarrieß, 2009) to map LFG f-structures to meaning representations. For the reverse direction, the meaning representations are mapped to f-structures which can then be mapped to surface strings by the XLE generator (Zarrieß and Kuhn, 2010).

For the final realisation ranking step in both pipelines, we used SVMrank, a Support Vector Machine-based learning tool (Joachims, 1996). The ranking step is thus technically independent from the LFG-based component. However, the grammar is used to produce the training data, pairs of corpus sentences and the possible alternations.

The two pipelines allow us to vary the degree to which the generation input is underspecified. An f-structure abstracts away from word order, i.e. the candidate set will contain just word order alternations. In the semantic input, syntactic function and voice are underspecified, so a larger set of surface realisation candidates is generated. Figure 2 illustrates the two representation levels for an active and

[2]Compare the bidirectional competition set-up in some Optimality-Theoretic work, e.g., (Kuhn, 2003).

[3]The choice among alternative f-structures is done with a discriminative model (Forst, 2007).

a passive sentence. The subject of the passive and the object of the active f-structure are mapped to the same role (patient) in the meaning representation.

3.1 Issues with "naive" underspecification

In order to create an underspecified voice representation that does indeed leave open the realisation options available to the speaker/writer, it is often not sufficient to remove just the syntactic function information. For instance, the subject of the active sentence (2) is an arbitrary reference pronoun *man "one"* which cannot be used as an oblique agent in a passive, sentence (2-b) is ungrammatical.

(2) a. Man hat den Kanzler gesehen.
One has the chancellor seen.
b. *Der Kanzler wurde von man gesehen.
The chancellor was by one seen.

So, when combined with the grammar, the meaning representation for (2) in Figure 2 contains implicit information about the voice of the original corpus sentence; the candidate set will not include any passive realisations. However, a passive realisation without the oblique agent in the *by*-phrase, as in Example (3), is a very natural variant.

(3) Der Kanzler wurde gesehen.
The chancellor was seen.

The reverse situation arises frequently too: passive sentences where the agent role is not overtly realised. Given the standard, "analysis-oriented" meaning representation for Sentence (4) in Figure 2, the realiser will not generate an active realisation since the agent role cannot be instantiated by any phrase in the grammar. However, depending on the exact context there are typically options for realising the subject phrase in an active with very little descriptive content.

Ideally, one would like to account for these phenomena in a meaning representation that underspecifies the lexicalisation of discourse referents, and also captures the reference of implicit arguments. Especially the latter task has hardly been addressed in NLP applications (but see Gerber and Chai (2010)). In order to work around that problem, we implemented some simple heuristics which underspecify the realisation of certain verb arguments. These rules define: 1. a set of pronouns (generic and neutral pronouns, universal quantifiers) that correspond to "trivial" agents in active and implicit agents in passive sentences; 2. a set of prepositional adjuncts in passive sentences that correspond to subjects in active sentence (e.g. causative and instrumental prepositions like *durch* "*by means of*"); 3. certain syntactic contexts where special underspecification devices are needed, e.g. coordinations or embeddings, see Zarrieß and Kuhn (2010) for examples. In the following, we will distinguish 1-role transitives where the agent is "trivial" or implicit from 2-role transitives with a non-implicit agent.

	Active	Passive
2-role trans.	71% (82%)	10% (2%)
1-role trans.	11% (0%)	8% (16%)

Table 1: Distribution of voices in SEM_h (SEM_n)

By means of the extended underspecification rules for voice, the sentences in (2) and (3) receive an identical meaning representation. As a result, our surface realiser can produce an active alternation for (3) and a passive alternation for (2). In the following, we will refer to the extended representations as SEM_h ("heuristic semantics"), and to the original representations as SEM_n ("naive semantics").

We are aware of the fact that these approximations introduce some noise into the data and do not always represent the underlying referents correctly. For instance, the implicit agent in a passive need not be "trivial" but can correspond to an actual discourse referent. However, we consider these heuristics as a first step towards capturing an important discourse function of the passive alternation, namely the deletion of the agent role. If we did not treat the passives with an implicit agent on a par with certain actives, we would have to ignore a major portion of the passives occurring in corpus data.

Table 1 summarises the distribution of the voices for the heuristic meaning representation SEM_h on the data-set we will introduce in Section 4, with the distribution for the naive representation SEM_n in parentheses.

4 Experimental Set-up

Data To obtain a sizable set of realistic corpus examples for our experiments on voice alternations, we created our own dataset of input sentences and representations, instead of building on treebank examples as Cahill et al. (2007) do. We extracted 19,905 sentences, all containing at least one transitive verb,

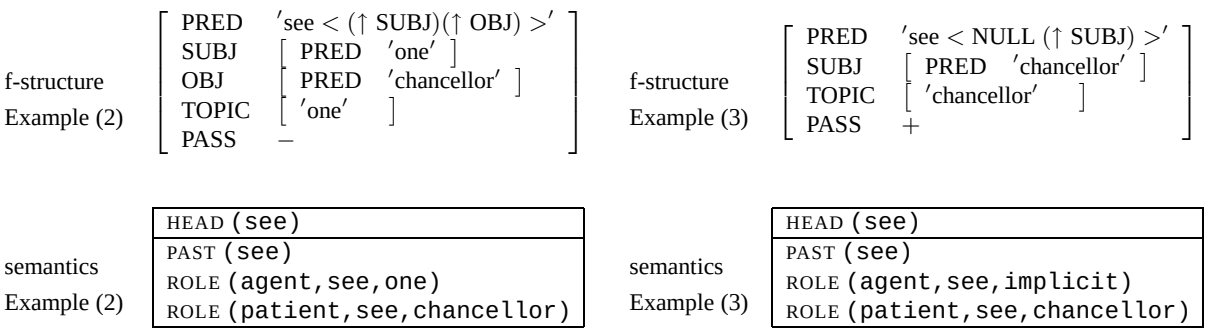

Figure 2: F-structure pair for passive-active alternation

from the HGC, a huge German corpus of newspaper text (204.5 million tokens). The sentences are automatically parsed with the German LFG grammar. The resulting f-structure parses are transferred to meaning representations and mapped back to f-structure charts. For our generation experiments, we only use those f-structure charts that the XLE generator can map back to a set of surface realisations. This results in a total of 1236 test sentences and 8044 sentences in our training set. The data loss is mostly due to the fact the XLE generator often fails on incomplete parses, and on very long sentences. Nevertheless, the average sentence length (17.28) and number of surface realisations (see Table 2) are higher than in Cahill et al. (2007).

Labelling For the training of our ranking model, we have to tell the learner how closely each surface realisation candidate resembles the original corpus sentence. We distinguish the rank categories: "1" identical to the corpus string, "2" identical to the corpus string ignoring punctuation, "3" small edit distance (< 4) to the corpus string ignoring punctuation, "4" different from the corpus sentence. In one of our experiments (Section 5.1), we used the rank category "5" to explicitly label the surface realisations derived from the alternation f-structure that does not correspond to the parse of the original corpus sentence. The intermediate rank categories "2" and "3" are useful since the grammar does not always regenerate the exact corpus string, see Cahill et al. (2007) for explanation.

Features The linguistic theories sketched in Section 2.2 correlate morphological, syntactic and semantic properties of constituents (or discourse referents) with their order and argument realisation. In our system, this correlation is modelled by a combination of linguistic properties that can be extracted from the f-structure or meaning representation and of the surface order that is read off the sentence string. Standard n-gram features are also used as features.[4] The feature model is built as follows: for every lemma in the f-structure, we extract a set of morphological properties (definiteness, person, pronominal status etc.), the voice of the verbal head, its syntactic and semantic role, and a set of informations status features following Cahill and Riester (2009). These properties are combined in two ways: a) Precedence features: relative order of properties in the surface string, e.g. "theme $<$ agent in passive", "1st person $<$ 3rd person"; b) "scale alignment" features (ScalAl.): combinations of voice and role properties with morphological properties, e.g. "subject is singular", "agent is 3rd person in active voice" (these are surface-independent, identical for each alternation candidate).

The model for which we present our results is based on sentence-internal features only; as Cahill and Riester (2009) showed, these feature carry a considerable amount of implicit information about the discourse context (e.g. in the shape of referring expressions). We also implemented a set of explicitly inter-sentential features, inspired by Centering Theory (Grosz et al., 1995). This model did not improve over the intra-sentential model.

Evaluation Measures In order to assess the general quality of our generation ranking models, we

[4]The language model is trained on the German data release for the 2009 ACL Workshop on Machine Translation shared task, 11,991,277 total sentences.

		FS	SEM$_n$	SEM$_h$
Avg. # strings		36.7	68.2	75.8
Random Match		16.98	10.72	7.28
LM	Match	15.45	15.04	11.89
	BLEU	0.68	0.68	0.65
	NIST	13.01	12.95	12.69
Ling. Model	Match	27.91	27.66	26.38
	BLEU	0.764	0.759	0.747
	NIST	13.18	13.14	13.01

Table 2: Evaluation of Experiment 1

use several standard measures: a) exact match: how often does the model select the original corpus sentence, b) BLEU: n-gram overlap between top-ranked and original sentence, c) NIST: modification of BLEU giving more weight to less frequent n-grams. Second, we are interested in the model's performance wrt. specific linguistic criteria. We report the following accuracies: d) Voice: how often does the model select a sentence realising the correct voice, e) Precedence: how often does the model generate the right order of the verb arguments (agent and patient), and f) Vorfeld: how often does the model correctly predict the verb arguments to appear in the sentence initial position before the finite verb, the so-called *Vorfeld*. See Sections 5.3 and 6 for a discussion of these measures.

5 Experiments

5.1 Exp. 1: Effect of Underspecified Input

We investigate the effect of the input's underspecification on a state-of-the-art surface realisation ranking model. This model implements the entire feature set described in Section 4 (it is further analysed in the subsequent experiments). We built 3 datasets from our alternation data: FS - candidates generated from the f-structure; SEM$_n$ - realisations from the naive meaning representations; SEM$_h$ - candidates from the heuristically underspecified meaning representation. Thus, we keep the set of original corpus sentences (=the target realisations) constant, but train and test the model on different candidate sets.

In Table 2, we compare the performance of the linguistically informed model described in Section 4 on the candidates sets against a random choice and a language model (LM) baseline. The differences in BLEU between the candidate sets and models are statistically significant.[5] In general, the linguistic model largely outperforms the LM and is less sensitive to the additional confusion introduced by the SEM$_h$ input. Its BLEU score and match accuracy decrease only slightly (though statistically significantly).

		FS	SEM$_n$	SEM$_h$	SEM$_n*$
All Trans.	Voice Acc.	100	98.06	91.05	97.59
	Voice Spec.	100	22.8	0	0
	Majority BL		82.4		98.1
2-role Trans.	Voice Acc.	100	97.7	91.8	97.59
	Voice Spec.	100	8.33	0	0
	Majority BL		88.5		98.1
1-role Trans.	Voice Acc.	100	100	90.0	-
	Voice Spec.	100	100	0	-
	Majority BL		53.9		-

Table 3: Accuracy of Voice Prediction by Ling. Model in Experiment 1

In Table 3, we report the performance of the linguistic model on the different candidate sets with respect to voice accuracy. Since the candidate sets differ in the proportion of items that underspecify the voice (see "Voice Spec." in Table 3), we also report the accuracy on the SEM$_n*$ test set, which is a subset of SEM$_n$ excluding the items where the voice is specified. Table 3 shows that the proportion of active realisations for the SEM$_n*$ input is very high, and the model does not outperform the majority baseline (which always selects active). In contrast, the SEM$_h$ model clearly outperforms the majority baseline.

Example (4) is a case from our development set where the SEM$_n$ model incorrectly predicts an active (4-a), and the SEM$_h$ correctly predicts a passive (4-b).

(4) a. 26 kostspielige Studien erwähnten die Finanzierung.
26 expensive studies mentioned the funding.

b. Die Finanzierung wurde von 26 kostspieligen Studien erwähnt.
The funding was by 26 expensive studies mentioned.

This prediction is according to the markedness hierarchy: the patient is singular and definite, the agent

[5] According to a bootstrap resampling test, $p < 0.05$

Features	Match	BLEU	Voice	Prec.	VF
Prec.	16.3	0.70	88.43	64.1	59.1
ScalAl.	10.4	0.64	90.37	58.9	56.3
Union	26.4	0.75	91.50	80.2	70.9

Table 4: Evaluation of Experiment 2

is plural and indefinite. Counterexamples are possible, but there is a clear statistical preference – which the model was able to pick up.

On the one hand, the rankers can cope surprisingly well with the additional realisations obtained from the meaning representations. According to the global sentence overlap measures, their quality is not seriously impaired. On the other hand, the design of the representations has a substantial effect on the prediction of the alternations. The SEM$_n$ does not seem to learn certain preferences because of the extremely imbalanced distribution in the input data. This confirms the hypothesis sketched in Section 3.1, according to which the degree of the input's underspecification can crucially change the behaviour of the ranking model.

5.2 Exp. 2: Word Order and Voice

We examine the impact of certain feature types on the prediction of the variation types in our data. We are particularly interested in the interaction of voice and word order (precedence) since linguistic theories (see Section 2.2) predict similar information-structural factors guiding their use, but usually do not consider them in conjunction.

In Table 4, we report the performance of ranking models trained on the different feature subsets introduced in Section 4. The union of the features corresponds to the model trained on SEM$_h$ in Experiment 1. At a very broad level, the results suggest that the precedence and the scale alignment features interact both in the prediction of voice and word order.

The most pronounced effect on voice accuracy can be seen when comparing the precedence model to the union model. Adding the surface-independent scale alignment features to the precedence features leads to a big improvement in the prediction of word order. This is not a trivial observation since a) the surface-independent features do not discriminate between the word orders and b) the precedence features are built from the same properties (see Section 4). Thus, the SVM learner discovers dependencies between relative precedence preferences and abstract properties of a verb argument which cannot be encoded in the precedence alone.

It is worth noting that the precedence features improve the voice prediction. This indicates that wherever the application context allows it, voice should not be specified at a stage prior to word order. Example (5) is taken from our development set, illustrating a case where the union model predicted the correct voice and word order (5-a), and the scale alignment model top-ranked the incorrect voice and word order. The active verb arguments in (5-b) are both case-ambigous and placed in the non-canonical order (object $<$ subject), so the semantic relation can be easily misunderstood. The passive in (5-a) is unambiguous since the agent is realised in a PP (and placed in the Vorfeld).

(5) a. Von den deutschen Medien wurden die Ausländer
By the German media were the foreigners
nur erwähnt, wenn es Zoff gab.
only mentioned, when there trouble was.

b. Wenn es Zoff gab, erwähnten die Ausländer
When there trouble was, mentioned the foreigners
nur die deutschen Medien.
only the German media.

Moreover, our results confirm Filippova and Strube (2007) who find that it is harder to predict the correct Vorfeld occupant in a German sentence, than to predict the relative order of the constituents.

5.3 Exp. 3: Capturing Flexible Variation

The previous experiment has shown that there is a certain inter-dependence between word order and voice. This experiment addresses this interaction by varying the way the training data for the ranker is labelled. We contrast two ways of labelling the sentences (see Section 4): a) all sentences that are not (nearly) identical to the reference sentence have the rank category "4", irrespective of their voice (referred to as unlabelled model), b) the sentences that do not realise the correct voice are ranked lower than sentences with the correct voice ("4" vs. "5"), referred to as labelled model. Intuitively, the latter way of labelling tells the ranker that all sentences in the incorrect voice are worse than all sentences in the correct voice, independent of the word order. Given the first labelling strategy, the ranker can decide in an unsupervised way which combinations of word order and voice are to be preferred.

Model	Match	BLEU	NIST	Top 1 Voice	Top 1 Prec.	Top 1 Prec.+Voice	Top 2 Prec.+Voice	Top 3 Prec.+Voice
Labelled, no LM	21.52	0.73	12.93	91.9	76.25	71.01	78.35	82.31
Unlabelled, no LM	26.83	0.75	13.01	91.5	80.19	74.51	84.28	88.59
Unlabeled + LM	27.35	0.75	13.08	91.5	79.6	73.92	79.74	82.89

Table 5: Evaluation of Experiment 3

In Table 5, it can be seen that the unlabelled model improves over the labelled on all the sentence overlap measures. The improvements are statistically significant. Moreover, we compare the n-best accuracies achieved by the models for the joint prediction of voice and argument order. The unlabelled model is very flexible with respect to the word order-voice interaction: the accuracy dramatically improves when looking at the top 3 sentences. Table 5 also reports the performance of an unlabelled model that additionally integrates LM scores. Surprisingly, these scores have a very small positive effect on the sentence overlap features and no positive effect on the voice and precedence accuracy. The n-best evaluations even suggest that the LM scores negatively impact the ranker: the accuracy for the top 3 sentences increases much less as compared to the model that does not integrate LM scores.[6]

The n-best performance of a realisation ranker is practically relevant for re-ranking applications such as Velldal (2008). We think that it is also conceptually interesting. Previous evaluation studies suggest that the original corpus sentence is not always the only optimal realisation of a given linguistic input (Cahill and Forst, 2010; Belz and Kow, 2010). Humans seem to have varying preferences for word order contrasts in certain contexts. The n-best evaluation could reflect the behaviour of a ranking model with respect to the range of variations encountered in real discourse. The pilot human evaluation in the next Section deals with this question.

6 Human Evaluation

Our experiment in Section 5.3 has shown that the accuracy of our linguistically informed ranking model dramatically increases when we consider the three best sentences rather than only the top-ranked sentence. This means that the model sometimes predicts almost equal naturalness for different voice realisations. Moreover, in the case of word order, we know from previous evaluation studies, that humans sometimes prefer different realisations than the original corpus sentences. This Section investigates agreement in human judgements of voice realisation.

Whereas previous studies in generation mainly used human evaluation to compare different systems, or to correlate human and automatic evaluations, our primary interest is the agreement or correlation between human rankings. In particular, we explore the hypothesis that this agreement is higher in certain contexts than in others. In order to select these contexts, we use the predictions made by our ranking model.

The questionnaire for our experiment comprised 24 items falling into 3 classes: a) items where the 3 best sentences predicted by the model have the same voice as the original sentence ("Correct"), b) items where the 3 top-ranked sentences realise different voices ("Mixed"), c) items where the model predicted the incorrect voice in all 3 top sentences ("False"). Each item is composed of the original sentence, the 3 top-ranked sentences (if not identical to the corpus sentence) and 2 further sentences such that each item contains different voices. For each item, we presented the previous context sentence.

The experiment was completed by 8 participants, all native speakers of German, 5 had a linguistic background. The participants were asked to rank each sentence on a scale from 1-6 according to its naturalness and plausibility in the given context. The participants were explicitly allowed to use the same rank for sentences they find equally natural. The participants made heavy use of this option: out of the 192 annotated items, only 8 are ranked such that no two sentences have the same rank.

We compare the human judgements by correlat-

[6](Nakanishi et al., 2005) also note a negative effect of including LM scores in their model, pointing out that the LM was not trained on enough data. The corpus used for training our LM might also have been too small or distinct in genre.

ing them with Spearman's ρ. This measure is considered appropriate for graded annotation tasks in general (Erk and McCarthy, 2009), and has also been used for analysing human realisation rankings (Velldal, 2008; Cahill and Forst, 2010). We normalise the ranks according to the procedure in Velldal (2008). In Table 6, we report the correlations obtained from averaging over all pairwise correlations between the participants and the correlations restricted to the item and sentence classes. We used bootstrap re-sampling on the pairwise correlations to test that the correlations on the different item classes significantly differ from each other.

The correlations in Table 6 suggest that the agreement between annotators is highest on the false items, and lowest on the mixed items. Humans tended to give the best rank to the original sentence more often on the false items (91%) than on the others. Moreover, the agreement is generally higher on the sentences realising the correct voice.

These results seem to confirm our hypothesis that the general level of agreement between humans differs depending on the context. However, one has to be careful in relating the effects in our data solely to voice preferences. Since the sentences were chosen automatically, some examples contain very unnatural word orders that probably guided the annotators' decisions more than the voice. This is illustrated by Example (6) showing two passive sentences from our questionnaire which differ only in the position of the adverb *besser "better"*. Sentence (6-a) is completely implausible for a native speaker of German, whereas Sentence (6-b) sounds very natural.

(6) a. Durch das neue Gesetz sollen **besser**
By the new law should better
Eigenheimbesitzer geschützt werden.
house owners protected be.

b. Durch das neue Gesetz sollen Eigenheimbesitzer
By the new law should house owners
besser geschützt werden.
better protected be.

This observation brings us back to our initial point that the surface realisation task is especially challenging due to the interaction of a range of semantic and discourse phenomena. Obviously, this interaction makes it difficult to single out preferences for a specific alternation type. Future work will have to establish how this problem should be dealt with in the design of human evaluation experiments.

	Items			
	All	Correct	Mixed	False
"All" sent.	0.58	0.6	0.54	0.62
"Correct" sent.	0.64	0.63	0.56	0.72
"False" sent.	0.47	0.57	0.48	0.44
Top-ranked corpus sent.	84%	78%	83%	91%

Table 6: Human Evaluation

7 Conclusion

We have presented a grammar-based generation architecture which implements the surface realisation of meaning representations abstracting from voice and word order. In order to be able to study voice alternations in a variety of contexts, we designed heuristic underspecification rules which establish, for instance, the alternation relation between an active with a generic agent and a passive that does not overtly realise the agent. This strategy leads to a better balanced distribution of the alternations in the training data, such that our linguistically informed generation ranking model achieves high BLEU scores and accurately predicts active and passive. In future work, we will extend our experiments to a wider range of alternations and try to capture inter-sentential context more explicitly. Moreover, it would be interesting to carry over our methodology to a purely statistical linearisation system where the relation between an input representation and a set of candidate realisations is not so clearly defined as in a grammar-based system.

Our study also addressed the interaction of different linguistic variation types, i.e. word order and voice, by looking at different types of linguistic features and exploring different ways of labelling the training data. However, our SVM-based learning framework is not well-suited to directly assess the correlation between a certain feature (or feature combination) and the occurrence of an alternation. Therefore, it would be interesting to relate our work to the techniques used in theoretical papers, e.g. (Bresnan et al., 2007), where these correlations are analysed more directly.

References

Judith Aissen. 1999. Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory*, 17(4):673–711.

Judith Aissen. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory*, 21:435–483.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*.

Anja Belz, Mike White, Josef van Genabith, Deirdre Hogan, and Amanda Stent. 2010. Finding common ground: Towards a surface realisation shared task. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*.

Anja Belz. 2005. Statistical generation: Three methods compared and evaluated. In *Proceedings of Tenth European Workshop on Natural Language Generation (ENLG-05)*, pages 15–23.

Bernd Bohnet, Leo Wanner, Simon Mill, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.

Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft Constraints Mirror Hard Constraints: Voice and Person in English and Lummi. In *Proceedings of the LFG '01 Conference*.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science.

Aoife Cahill and Martin Forst. 2010. Human Evaluation of a German Surface Realisation Ranker. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 112 – 120, Athens, Greece. Association for Computational Linguistics.

Aoife Cahill and Arndt Riester. 2009. Incorporating Information Status into Generation Ranking. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 817–825, Suntec, Singapore, August. Association for Computational Linguistics.

Aoife Cahill, Martin Forst, and Christian Rohrer. 2007. Stochastic realisation ranking for a free word order language. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 17–24, Saarbrücken, Germany, June. DFKI GmbH. Document D-07-01.

Dick Crouch and Tracy Holloway King. 2006. Semantics via F-Structure Rewriting. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG06 Conference*.

Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2006. XLE Documentation. Technical report, Palo Alto Research Center, CA.

Katrin Erk and Diana McCarthy. 2009. Graded Word Sense Assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440 – 449, Singapore.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 07)*, Prague, Czech Republic.

Katja Filippova and Michael Strube. 2009. Tree linearization in English: Improving language model based approaches. In *Companion Volume to the Proceedings of Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 09, short).*, Boulder, Colorado.

Martin Forst. 2007. Filling Statistics with Linguistics – Property Design for the Disambiguation of German LFG Parses. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Matthew Gerber and Joyce Chai. 2010. Beyond nombank: A study of implicit argumentation for nominal predicates. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Barbara J. Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Thorsten Joachims. 1996. Training linear svms in linear time. In M. Butt and T. H. King, editors, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, CSLI Proceedings Online.

Jonas Kuhn. 2003. *Optimality-Theoretic Syntax—A Declarative Approach*. CSLI Publications, Stanford, CA.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the ACL/COLING-98*, pages 704–710, Montreal, Quebec.

Tomasz Marciniak and Michael Strube. 2005. Using an annotated corpus as a knowledge source for language generation. In *Proceedings of Workshop on Using Corpora for Natural Language Generation*, pages 19–24, Birmingham, UK.

Hiroko Nakanishi, Yusuke Miyao, and Junichi Tsujii. 2005. Probabilistic models for disambiguation of an

HPSG-based chart generator. In *Proceedings of IWPT 2005*.

Adwait Ratnaparkhi. 2000. Trainable methods for surface natural language generation. In *Proceedings of NAACL 2000*, pages 194–201, Seattle, WA.

Stefan Riezler, Detlef Prescher, Jonas Kuhn, and Mark Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00), Hong Kong*, pages 480–487.

Stefan Riezler, Dick Crouch, Ron Kaplan, Tracy King, John Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Pennsylvania, Philadelphia*.

Eric K. Ringger, Michael Gamon, Robert C. Moore, David Rojas, Martine Smets, and Simon Corston-Oliver. 2004. Linguistically Informed Statistical Models of Constituent Structure for Ordering in Sentence Realization. In *Proceedings of the 2004 International Conference on Computational Linguistics*, Geneva, Switzerland.

Christian Rohrer and Martin Forst. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of LREC-2006*.

Erik Velldal and Stephan Oepen. 2006. Statistical ranking in tactical generation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Erik Velldal. 2008. *Empirical Realization Ranking*. Ph.D. thesis, University of Oslo, Department of Informatics.

Sina Zarrieß and Jonas Kuhn. 2010. Reversing F-structure Rewriting for Generation from Meaning Representations. In *Proceedings of the LFG10 Conference*, Ottawa.

Sina Zarrieß. 2009. Developing German Semantics on the basis of Parallel LFG Grammars. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks (GEAF 2009)*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.