Human Evaluation of a German Surface Realisation Ranker

Aoife Cahill
Institut für Maschinelle Sprachverarbeitung (IMS)
University of Stuttgart
70174 Stuttgart, Germany
aoife.cahill@ims.uni-stuttgart.de

Martin Forst
Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304, USA
mforst@parc.com

Abstract

In this paper we present a human-based evaluation of surface realisation alternatives. We examine the relative rankings of naturally occurring corpus sentences and automatically generated strings chosen by statistical models (language model, log-linear model), as well as the naturalness of the strings chosen by the log-linear model. We also investigate to what extent preceding context has an effect on choice. We show that native speakers do accept quite some variation in word order, but there are also clearly factors that make certain realisation alternatives more natural.

1 Introduction

An important component of research on surface realisation (the task of generating strings for a given abstract representation) is evaluation, especially if we want to be able to compare across systems. There is consensus that exact match with respect to an actually observed corpus sentence is too strict a metric and that BLEU score measured against corpus sentences can only give a rough impression of the quality of the system output. It is unclear, however, what kind of metric would be most suitable for the evaluation of string realisations, so that, as a result, there have been a range of automatic metrics applied including *inter alia* exact match, string edit distance, NIST SSA, BLEU, NIST, ROUGE, generation string accuracy, generation tree accuracy, word accuracy (Bangalore et al., 2000; Callaway, 2003; Nakanishi et al., 2005; Velldal and Oepen, 2006; Belz and Reiter, 2006).

It is not always clear how appropriate these metrics are, especially at the level of individual sentences. Using automatic evaluation metrics cannot be avoided, but ideally, a metric for the evaluation of realisation rankers would rank alternative realisations in the same way as native speakers of the language for which the surface realisation system is developed, and not only globally, but also at the level of individual sentences.

Another major consideration in evaluation is what to take as the gold standard. The easiest option is to take the original corpus string that was used to produce the abstract representation from which we generate. However, there may well be other realisations of the same input that are as suitable in the given context. Reiter and Sripada (2002) argue that while we should take advantage of large corpora in NLG, we also need to take care that we do not introduce errors by learning from incorrect data present in corpora.

In order to better understand what makes good evaluation data (and metrics), we designed and implemented an experiment in which human judges evaluated German string realisations. The main aims of this experiment were: (i) to establish how much variation in German word order is acceptable for human judges, (ii) to find an automatic evaluation metric that mirrors the findings of the human evaluation, (iii) to provide detailed feedback for the designers of the surface realisation ranking model and (iv) to establish what effect preceding context has on the choice of realisation. In this paper, we concentrate on points (i) and (iv).

The remainder of the paper is structured as follows: In Section 2 we outline the realisation ranking system that provided the data for the experiment. In Section 3 we outline the design of the experiment and in Section 4 we present our findings. In Section 5 we relate this to other work and finally we conclude in Section 6.

2 A Realisation Ranking System for German

We take the realisation ranking system for German described in Cahill et al. (2007) and present the output to human judges. One goal of this series of experiments is to examine whether the results

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 112–120,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

based on automatic evaluation metrics published in that paper are confirmed in an evaluation by humans. Another goal is to collect data that will allow us and other researchers[1] to explore more fine-grained and reliable automatic evaluation metrics for realisation ranking.

The system presented by Cahill et al. (2007) ranks the strings generated by a hand-crafted broad-coverage Lexical Functional Grammar (Bresnan, 2001) for German (Rohrer and Forst, 2006) on the basis of a given input f-structure. In these experiments, we use f-structures from their held-out and test sets, of which 96% can be associated with surface realisations by the grammar. F-structures are attribute-value matrices representing grammatical functions and morphosyntactic features; roughly speaking, they are predicate-argument structures. In LFG, f-structures are assumed to be a crosslinguistically relatively parallel syntactic representation level, alongside the more surface-oriented c-structures, which are context-free trees. Figure 1 shows the f-structure[2] associated with TIGER Corpus sentence 8609, glossed in (1), as well as the 4 string realisations that the German LFG generates from this f-structure. The LFG is reversible, i.e. the same grammar is used for parsing as for generation. It is a hand-crafted grammar, and has been carefully constructed to only parse (and therefore generate) grammatical strings.[3]

(1)	Williams	war	in	der	britischen	Politik	äußerst
	Williams	was	in	the	British	politics	extremely
	umstritten.						
	controversial.						

‘Williams was extremely controversial in British politics.’

The ranker consists of a log-linear model that is based on linguistically informed structural features as well as a trigram language model, whose score is integrated into the model simply as an additional feature. The log-linear model is trained on corpus data, in this case sentences from the TIGER Corpus (Brants et al., 2002), for which f-structures are available; the observed corpus sentences are considered as references whose probability is to be maximised during the training process.

The output of the realisation ranker is evaluated in terms of exact match and BLEU score, both measured against the actually observed corpus sentences. In addition to the figures achieved by the ranker, the corresponding figures achieved by the employed trigram language model on its own are given as a baseline, and the exact match figure of the best possible string selection is given as an upper bound.[4] We summarise these figures in Table 1.

	Exact Match	BLEU score
Language model	27%	0.7306
Log-linear model	37%	0.7939
Upper bound	62%	–

Table 1: Results achieved by trigram LM ranker and log-linear model ranker in Cahill et al. (2007)

By means of these figures, Cahill et al. (2007) show that a log-linear model based on structural features and a language model score performs considerably better realisation ranking than just a language model. In our experiments, presented in detail in the following section, we examine whether human judges confirm this and how natural and/or acceptable the selection performed by the realisation ranker under consideration is for German native speakers.

3 Experiment Design

The experiment was divided into three parts. Each part took between 30 and 45 minutes to complete, and participants were asked to leave some time (e.g. a week) between each part. In total, 24 participants completed the experiment. All were native German speakers (mostly from South-Western Germany) and almost all had a linguistic background. Table 2 gives a breakdown of the items in each part of the experiment.[5]

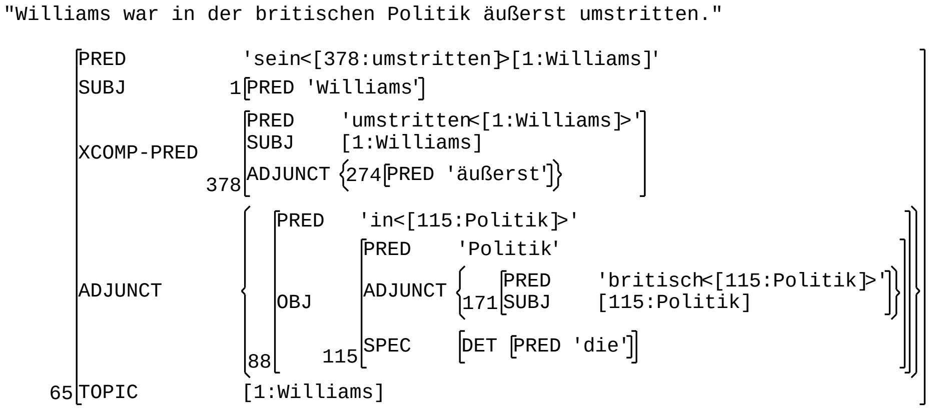
[1] The data is available for download from http://www.ims.uni-stuttgart.de/projekte/pargram/geneval/data/

[2] Note that only grammatical functions are displayed; morphosyntactic features are omitted due to space constraints. Also note that the discourse function TOPIC was ignored in generation.

[3] A ranking mechanism based on so-called optimality marks can lead to a certain “asymmetry” between parsing and generation in the sense that not all sentences that can be associated with a certain f-structure are necessarily generated from this same f-structure. E.g. the sentence *Williams war äußerst umstritten in der britischen Politik.* can be parsed into the f-structure in Figure 1, but it is not generated because an optimality mark penalizes the extraposition of PPs to the right of a clause. Only few optimality marks were used in the process of generating the data for our experiments, so that the bias they introduce should not be too noticeable.

[4] The observed corpus sentence can be (re)generated from the corresponding f-structure for only 62% of the sentences used, usually because of differences in punctuation. Hence this exact match upper bound. An upper bound in terms of BLEU score cannot be computed because BLEU score is computed on entire corpora rather than individual sentences.

[5] Experiments 3a and 3b contained the same items as experiments 1a and 1b.


"Williams war in der britischen Politik äußerst umstritten."
PRED 'sein<[378:umstritten]>[1:Williams]'
SUBJ 1[PRED 'Williams']
XCOMP-PRED
PRED 'umstritten<[1:Williams]>'
SUBJ [1:Williams]
ADJUNCT {274[PRED 'äußerst']}
378
ADJUNCT
PRED 'in<[115:Politik]>'
OBJ
PRED 'Politik'
ADJUNCT
PRED 'britisch<[115:Politik]>'
171 SUBJ [115:Politik]
115 SPEC [DET [PRED 'die']]
88
65 TOPIC [1:Williams]

```
Williams war in der britischen Politik äußerst umstritten.
In der britischen Politik war Williams äußerst umstritten.
Äußerst umstritten war Williams in der britischen Politik.
Äußerst umstritten war in der britischen Politik Williams.
```

Figure 1: F-structure associated with (1) and strings generated from it.

	Exp 1a	Exp 1b	Exp 2
Num. items	44	52	41
Avg. sent length	14.4	12.1	9.4

Table 2: Statistics for each experiment part

3.1 Part 1

The aim of part 1 of the experiment was twofold. First, to identify the relative rankings of the systems evaluated in Cahill et al. (2007) according to the human judges, and second to evaluate the quality of the strings as chosen by the log-linear model of Cahill et al. (2007). To these ends, part 1 was further subdivided into two tasks: 1a and b.

Task 1a: During the first task, participants were presented with alternative realisations for an input f-structure (but not shown the original f-structure) and asked to rank them in order of how natural sounding they were, 1 being the best and 3 being the worst.[6] Each item contained three alternatives, (i) the original string found in TIGER, (ii) the string chosen as most likely by the trigram language model, and (iii) the string chosen as most likely by the log-linear model. Only items where each system chose a different alternative were chosen from the evaluation data of Cahill et al. (2007). The three alternatives were presented in random order for each item, and the items were presented in random order for each participant. Some items were presented randomly to participants more than once as a sanity check, and in total for Part 1a, participants made 52 ranking judgements on 44 items. Figure 2 shows a screen shot of what the participant was presented with for this task.

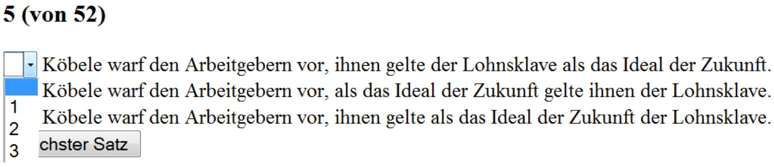
Task 1b: In the second task of part 1, participants were presented with the string chosen by the log-linear model as being the most likely and asked to evaluate it on a scale from 1 to 5 on how natural sounding it was, 1 being very unnatural or marked and 5 being completely natural. Figure 3 shows a screen shot of what the participant saw during the experiment. Again some random items were presented to the participant more than once, and the items themselves were presented in random order. In total, the participants made 58 judgements on 52 items.

3.2 Part 2

In the second part of the experiment, participants were presented between 4 and 8 alternative surface realisations for an input f-structure, as well as some preceding context. This preceding context was automatically determined using information from the export release of the TIGER treebank and was not hand-checked for relevance.[7] The participants were then asked to choose the realisation that they felt fit best given the preceding sentences.

[6] Joint rankings were not allowed, i.e. the participants were forced to make strict ranking decisions, and in hindsight this may have introduced some noise into the data.

[7] The export release of the TIGER treebank includes an article ID for each sentence. Unfortunately, this is not completely reliable for determining relevant context, since an article can also contain several short news snippets which are completely unrelated. Paragraph boundaries are not marked. This leads to some noise, which unfortunately is difficult to measure objectively



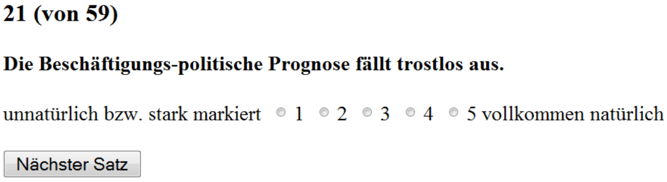
5 (von 52)

Köbele warf den Arbeitgebern vor, ihnen gelte der Lohnsklave als das Ideal der Zukunft.

Köbele warf den Arbeitgebern vor, als das Ideal der Zukunft gelte ihnen der Lohnsklave.

Köbele warf den Arbeitgebern vor, ihnen gelte als das Ideal der Zukunft der Lohnsklave.

Figure 2: Screenshot of Part 1a of the Experiment



21 (von 59)

Die Beschäftigungs-politische Prognose fällt trostlos aus.

unnatürlich bzw. stark markiert ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 vollkommen natürlich

Figure 3: Screenshot of Part 1b of the Experiment

	Total			Average
	Rank 1	Rank 2	Rank 3	Rank
Original String	817	366	65	1.40
LL String	303	593	352	2.04
LM String	128	289	831	2.56

Table 3: Task 1a: Ranks for each system

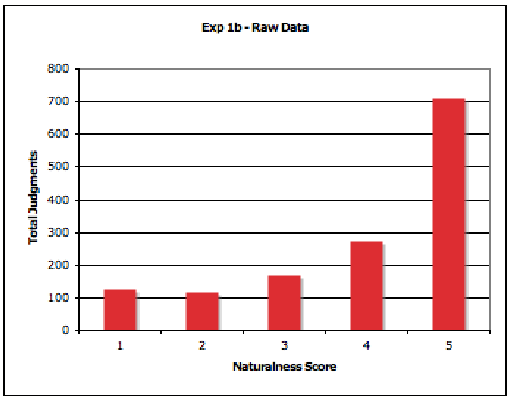


Figure 5: Task 1b: Naturalness scores for strings chosen by log-linear model, 1=worst

The items were presented in random order, and the list of alternatives were presented in random order to each participant. Some items were randomly presented more than once, resulting in 50 judgements on 41 items. Figure 4 shows a screen shot of what the participant saw.

3.3 Part 3

Part 3 of the experiment was identical to Part 1, except that now, rather than the participants being presented with sentences in isolation, they were given some preceding context. The context was determined automatically, in the same way as in Part 2. The items themselves were the same as in Part 1. The aim of this part of the experiment was to see what effect preceding context had on judgements.

4 Results

In this section we present the result and analysis of the experiments outlined above.

4.1 How good were the strings?

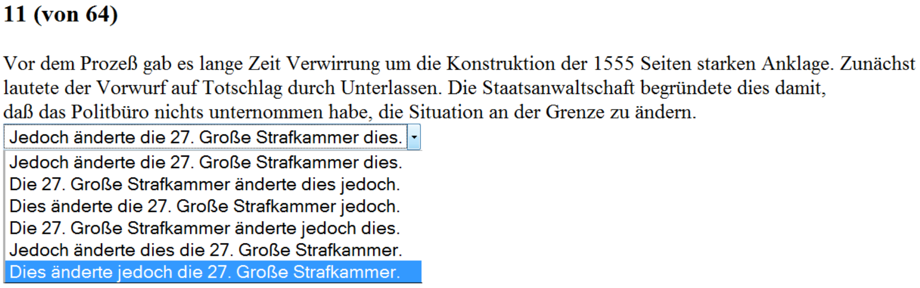
The data collected in Experiment 1a showed the overall human relative ranking of the three systems. We calculate the total numbers of each rank for each system. Table 3 summarises the results. The original string is the string found in the TIGER Corpus, the LM String is the string chosen as being most likely by the trigram language model and the LL String is the string chosen as being most likely by the log-linear model.

Table 3 confirms the overall relative rankings of the three systems as determined using BLEU scores. The original TIGER strings are ranked best (average 1.4), the strings chosen by the log-linear model are ranked better than the strings chosen by the language model (average 2.65 vs 2.04).

In Experiment 1b, the aim was to find out how acceptable the strings chosen by the log-linear model were, although they were not the same as the original string. Figure 5 summarises the data. The graph shows that the majority of strings chosen by the log-linear model ranked very highly on the naturalness scale.

4.2 Did the human judges agree with the original authors?

In Experiment 2, the aim was to find out how often the human judges chose the same string as the original author (given alternatives generated by the LFG grammar). Most items had between 4 and 6 alternative strings. In 70% of all items, the human judges chose the same string as the original author. However, the remaining 30% of the time, the human judges picked an alternative as being the



11 (von 64)

Vor dem Prozeß gab es lange Zeit Verwirrung um die Konstruktion der 1555 Seiten starken Anklage. Zunächst lautete der Vorwurf auf Totschlag durch Unterlassen. Die Staatsanwaltschaft begründete dies damit, daß das Politbüro nichts unternommen habe, die Situation an der Grenze zu ändern.

Jedoch änderte die 27. Große Strafkammer dies.

Jedoch änderte die 27. Große Strafkammer dies.
Die 27. Große Strafkammer änderte dies jedoch.
Dies änderte die 27. Große Strafkammer jedoch.
Die 27. Große Strafkammer änderte jedoch dies.
Jedoch änderte dies die 27. Große Strafkammer.
Dies änderte jedoch die 27. Große Strafkammer.

Figure 4: Screenshot of Part 2 of the Experiment

most fitting in the given context.[8] This suggests that there is quite some variation in what native German speakers will accept, but that this variation is by no means random, as indicated by 70% of choices being the same string as the original author's.

Figure 6 shows for each bin of possible alternatives, the percentage of items with a given number of choices made. For example, for the items with 4 possible alternatives, over 70% of the time, the judges chose between only 2 of them. For the items with 5 possible alternatives, in 10% of those items the human judges chose only 1 of those alternatives; in 30% of cases, the human judges all chose the same 2 solutions, and for the remaining 60% they chose between only 3 of the 5 possible alternatives. These figures indicate that although judges could not always agree on one best string, often they were only choosing between 2 or 3 of the possible alternatives. This suggests that, on the one hand, native speakers do accept quite some variation, but that, on the other hand, there are clearly factors that make certain realisation alternatives more preferable than others.

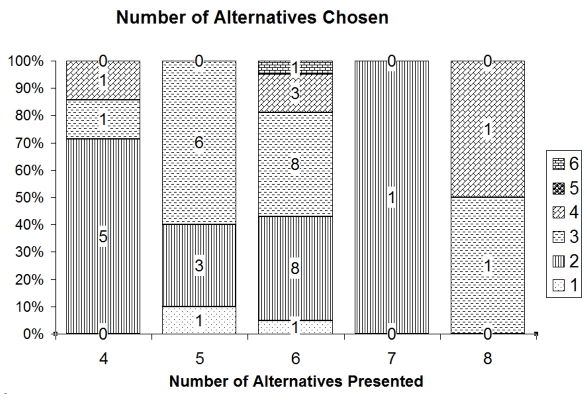


Figure 6: Exp 2: Number of Alternatives Chosen

The graph in Figure 6 shows that only in two cases did the human judges choose from among all possible alternatives. In one case, there were 4 possible alternatives and in the other 6. The original sentence that had 4 alternatives is given in (2). The four alternatives that participants were asked to choose from are given in Table 4, with the frequency of each choice. The original sentence that had 6 alternatives is given in (3). The six alternatives generated by the grammar and the frequencies with which they were chosen is given in Table 5.

(2) Die Brandursache blieb zunächst unbekannt.
The cause of fire remained initially unknown.
'The cause of the fire remained unknown initially.'

Alternative	Freq.
Zunächst blieb die Brandursache unbekannt.	2
Die Brandursache blieb zunächst unbekannt.	24
Unbekannt blieb die Brandursache zunächst.	1
Unbekannt blieb zunächst die Brandursache.	1

Table 4: The 4 alternatives given by the grammar for (2) and their frequencies

Tables 4 and 5 tell different stories. On the one hand, although each of the 4 alternatives was chosen at least once from Table 4, there is a clear preference for one string (and this is also the original string from the TIGER Corpus). On the other hand, there is no clear preference[9] for any one of the alternatives in Table 5, and, in fact, the alternative that was selected most frequently by the participants is not the original string. Interestingly, out of the 41 items presented to participants, the original string was chosen by the majority of participants in 36 cases. Again, this confirms the hypothesis that there is a certain amount of acceptable variation for native speakers but there are clear preferences for certain strings over others.

[8] Recall that almost all strings presented to the judges were grammatical.

[9] Although it is clear that alternative 2 is dispreferred.

(3) Die Unternehmensgruppe Tengelmann fördert mit einem sechsstelligen Betrag die Arbeit im brandenburgischen
The group of companies Tengelmann assists with a 6-figure sum the work in of-Brandenburg
Biosphärenreservat Schorfheide.
biosphere reserve Schorfheide.

'The Tengelmann group of companies is supporting the work at the biosphere reserve in Schorfheide, Brandenburg, with a 6-figure sum.'

Alternative	Freq.
Mit einem sechsstelligen Betrag fördert die Unternehmensgruppe Tengelmann die Arbeit im brandenburgischen Biosphärenreservat Schorfheide.	7
Mit einem sechsstelligen Betrag fördert die Arbeit im brandenburgischen Biosphärenreservat Schorfheide die Unternehmensgruppe Tengelmann.	1
Die Arbeit im brandenburgischen Biosphärenreservat Schorfheide fördert die Unternehmensgruppe Tengelmann mit einem sechsstelligen Betrag.	4
Die Arbeit im brandenburgischen Biosphärenreservat Schorfheide fördert mit einem sechsstelligen Betrag die Unternehmensgruppe Tengelmann.	5
Die Unternehmensgruppe Tengelmann fördert die Arbeit im brandenburgischen Biosphärenreservat Schorfheide mit einem sechsstelligen Betrag.	5
Die Unternehmensgruppe Tengelmann fördert mit einem sechsstelligen Betrag die Arbeit im brandenburgischen Biosphärenreservat Schorfheide.	5

Table 5: The 6 alternatives given by the grammar for (3) and their frequencies

4.3 Effects of context

As explained in Section 3.1, Part 3 of our experiment was identical to Part 1, except that the participants could see some preceding context. The aim of this part was to investigate to what extent discourse factors influence the way in which human judges evaluate the output of the realisation ranker. In Task 3a, we expected the original strings to be ranked (even) higher in context than out of context; consequently, the ranks of the realisations selected by the log-linear and the language model would have to go down. With respect to Task 3b, we had no particular expectation, but were just interested in seeing whether some preceding context would affect the evaluation results for the strings selected as most probable by the log-linear model ranker in any way.

Table 6 summarises the results of Task 3a. It shows that, at least overall, our expectation that the original corpus sentences would be ranked higher within context than out of context was not borne out. Actually, they were ranked a bit lower than they were when presented in isolation, and the only realisations that are ranked slightly higher overall are the ones selected by the trigram LM.

	Total			Average
	Rank 1	Rank 2	Rank 3	Rank
Original String	810	365	71	1.41
	(-7)	(-1)	(+6)	(+0.01)
LL String	274	615	357	2.07
	(-29)	(+22)	(+5)	(+0.03)
LM String	162	266	818	2.53
	(+34)	(-23)	(-13)	(-0.03)

Table 6: Task 3a: Ranks for each system (compared to ranks in Task 1a)

The overall results of Task 3b are presented in Figure 7. Interestingly, although we did not expect any particular effect of preceding context on the way the participants would rate the realisations selected by the log-linear model, the naturalness scores were higher in the condition with context (Task 3b) than in the one without context (Task 1b). One explanation might be that sentences in some sort of default order are generally rated higher in context than out of context, simply because the context makes sentences less surprising.

Since, contrary to our expectations, we could not detect a clear effect of context in the overall results of Task 3a, we investigated how the average ranks of the three alternatives presented for individual items differ between Task 1a and Task 3a. An example of an original corpus sentence which many participants ranked higher in context than in isolation is given in (4a.). The realisations selected by the the log-linear model and the trigram LM are given in (4b.) and (4c.) respectively, and the context shown to the participants is given above these alternatives. We believe that the context has this effect because it prepares the reader for the structure with the sentence-initial predicative participle *entscheidend*; usually, these elements appear rather in clause-final position.

In contrast, (5a) is an example of a corpus

(4) -2 Betroffen sind die Antibabypillen Femovan, Lovelle, [...] und Dimirel.
Concerned are the contraceptive pills Femovan, Lovelle, [...], and Dimirel.

-1 Das Bundesinstitut schließt nicht aus, daß sich die Thrombose-Warnung als grundlos erweisen könnte.
The federal institute excludes not that the thrombosis warning as unfounded turn out could.

a. Entscheidend sei die [...] abschließende Bewertung, sagte Jürgen Beckmann vom Institut dem ZDF.
Decisive is the [...] final evaluation, said Jürgen Beckmann of the institute the ZDF.

b. Die [...] abschließende Bewertung sei entscheidend, sagte Jürgen Beckmann vom Institut dem ZDF.

c. Die [...] abschließende Bewertung sei entscheidend, sagte dem ZDF Jürgen Beckmann vom Institut.

(5) -2 Im konkreten Fall darf der Kurde allerdings trotz der Entscheidung der Bundesrichter nicht in die
In the concrete case may the Kurd however despite the decision of the federal judges not to the
Türkei abgeschoben werden, weil ihm dort nach den Feststellungen der Vorinstanz
Turkey deported be because him there according to the conclusions of the court of lower instance
politische Verfolgung droht.
political persecution threatens.

-1 Es besteht Abschiebeschutz nach dem Ausländergesetz.
It exists deportation protection according to the foreigner law.

a. Der 9. Senat [...] äußerte sich in seiner Entscheidung nicht zur Verfassungsgemäßheit der
The 9th senate [...] expressed itself in its decision not to the constitutionality of the
Drittstaatenregelung.
third-country rule.

b. In seiner Entscheidung äußerte sich der 9. Senat [...] nicht zur Verfassungsgemäßheit der Drittstaatenregelung.

c. Der 9. Senat [...] äußerte sich in seiner Entscheidung zur Verfassungsgemäßheit der Drittstaatenregelung nicht.

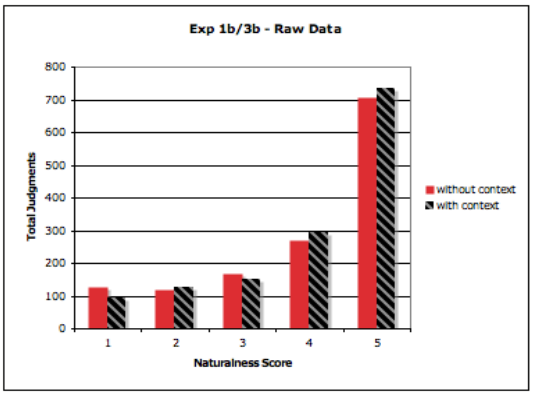


Figure 7: Tasks 1b and 3b: Naturalness scores for strings chosen by log-linear model, presented without and with context

sentence which our participants tended to rank lower in context than in isolation. Actually, the human judges preferred the realisation selected by the trigram LM to the original sentence and the realisation chosen by the log-linear model in both conditions, but this preference was even reinforced when context was available. One explanation might be that the two preceding sentences are precisely about the decision to which the initial phrase of variant (5b) refers, which ensures a smooth flow of the discourse.

4.4 Inter-Annotator Agreement

We measure two types of annotator agreement. First we measure how well each annotator agrees with him/herself. This is done by evaluating what percentage of the time an annotator made the same choice when presented with the same item choices (recall that as described in Section 3, a number of items were presented randomly more than once to each participant). The results are given in Table 7. The results show that in between 70% and 74% of cases, judges make the same decision when presented with the same data. We found this to be a surprisingly low number and think that it is most likely due to the acceptable variation in word order for speakers. Another measure of agreement is how well the individual participants agree with each other. In order to establish this, we calculate an average Spearman's correlation coefficient (non-parametric Pearson's correlation coefficient) between each participant for each experiment. The results are summarised in Table 8. Although these figures indicate a high level of inter-annotator agreement, more tests are required to establish exactly what these figures mean for each experiment.

5 Related Work

The work that is most closely related to what is presented in this paper is that of Velldal (2008). In

Experiment	Agreement (%)
Part 1a	77.43
Part 1b	71.05
Part 2	74.32
Part 3a	72.63
Part 3b	70.89

Table 7: How often did a participant make the same choice?

Experiment	Spearman coefficient
Part 1a	0.62
Part 1b	0.60
Part 2	0.58
Part 3a	0.61
Part 3b	0.51

Table 8: Inter-Annotator Agreement for each experiment

his thesis several models of realisation ranking are presented and evaluated against the original corpus text. Chapter 8 describes a small human-based experiment, where 7 native English speakers rank the output of 4 systems. One system is the original text, another is a randomly chosen baseline, another is a string chosen by a log-linear model and the fourth is one chosen by a language model. Joint rankings were allowed. The results presented in Velldal (2008) mirror our findings in Experiments 1a and 3a, that native speakers rank the original strings higher than the log-linear model strings which are ranked higher than the language model strings. In both cases, the log-linear models include the language model score as a feature in the log-linear model. Nakanishi et al. (2005) report that they achieve the best BLEU scores when they do not include the language model score in their log-linear model, but they also admit that their language model was not trained on enough data.

Belz and Reiter (2006) carry out a comparison of automatic evaluation metrics against human domain experts and human non-experts in the domain of weather forecast statements. In their evaluations, the NIST score correlated more closely than BLEU or ROUGE to the human judgements. They conclude that more than 4 reference texts are needed for automatic evaluation of NLG systems.

6 Conclusion and Outlook to Future Work

In this paper, we have presented a human-based experiment to evaluate the output of a realisation ranking system for German. We evaluated the original corpus text, and strings chosen by a language model and a log-linear model. We found that, at a global level, the human judgements mirrored the relative rankings of the three system according to the BLEU score. In terms of naturalness, the strings chosen by the log-linear model were generally given 4 or 5, indicating that although the log-linear model might not choose the same string as the original author had written, the strings it was choosing were mostly very natural strings.

When presented with all alternatives generated by the grammar for a given input f-structure, the human judges chose the same string as the original author 70% of the time. In 5 out of 41 cases, the majority of judges chose a string other than the original string. These figures show that native speakers accept some variation in word order, and so caution should be exercised when using corpus-derived reference data. The observed acceptable variation was often linked to information structural considerations, and further experiments will be carried out to investigate this relationship between word order and information structure.

In examining the effect of preceding context, we found that overall context had very little effect. At the level of individual sentences, however, clear tendencies were observed, but there were some sentences which were judged better in context and others which were ranked lower. This again indicates that corpus-derived reference data should be used with caution.

An obvious next step is to examine how well automatic metrics correlate with the human judgements collected, not only at an individual sentence level, but also at a global level. This can be done using statistical techniques to correlate the human judgements with the scores from the automatic metrics. We will also examine the sentences that were consistently judged to be of poor quality, so that we can provide feedback to the developers of the log-linear model in terms of possible additional features for disambiguation.

Acknowledgments

We are extremely grateful to all of our participants for taking part in this experiment. This work was partly funded by the Collaborative Research Centre (SFB 732) at the University of Stuttgart.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the First International Natural Language Generation Conference (INLG2000)*, pages 1–8, Mitzpe Ramon, Israel.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell, Oxford.

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.

Charles Callaway. 2003. Evaluating Coverage for Large Symbolic NLG Grammars. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pages 811–817, Acapulco, Mexico.

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

Ehud Reiter and Somayajulu Sripada. 2002. Should Corpora Texts Be Gold Standards for NLG? In *Proceedings of INLG-02*, pages 97–104, Harriman, NY.

Christian Rohrer and Martin Forst. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2006)*, Genoa, Italy.

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.