Improving Metrics for Speech Translation

Claudio Paonessa and **Dominik Frefel** and **Manfred Vogel**
Institute for Data Science
University of Applied Sciences and Arts Northwestern Switzerland
Windisch, Switzerland
claudio.paonessa@fhnw.ch

Abstract

We introduce Parallel Paraphrasing ($\text{Para}_{\text{both}}$), an augmentation method for translation metrics making use of automatic paraphrasing of both the reference and hypothesis. This method counteracts the typically misleading results of speech translation metrics such as WER, CER, and BLEU if only a single reference is available. We introduce two new datasets explicitly created to measure the quality of metrics intended to be applied to Swiss German speech-to-text systems. Based on these datasets, we show that we are able to significantly improve the correlation with human quality perception if our method is applied to commonly used metrics.

1 Introduction

For most practical use cases of speech-to-text systems, the transcription is perfect if the sentences are grammatically correct and the semantic meaning is fully captured. For speech translation, the phrasing of the transcription is ambiguous, contrary to word-by-word transcriptions for cases where input and output languages match. This means that a single reference rarely adequately covers the valid output range. The widely applied metrics BLEU (Papineni et al., 2002), Word Error Rate (WER), and Character Error Rate (CER) fail to handle the occurring sentence ambiguity, if calculated based on single references per sample. Many additional references are needed for each sample to counteract this flaw, which are often not available due to the high cost and effort of collecting them. This leads to metric results often not agreeing with the human perception of the transcription quality. More generally, the ambiguous translation space and therefore the need for multiple references occurs for all translation tasks.

Swiss German is a family of dialects and a mostly spoken language. The language lacks standardized writing, the written form usually only appears in informal text messages. Therefore, Swiss speech-to-text systems transform Swiss German speech to Standard German text, i.e., speech translation with a very similar source and target language. Because Swiss German is different from Standard German regarding phonetics, vocabulary, morphology, and syntax, the system output tends not to match with the single reference in a test set. Table 1 contains examples where common metrics fail to handle the paraphrased but semantically correct hypothesis.

Table 1: Examples of semantically matching but paraphrased references and hypotheses with corresponding metric values. See Section 4.3 for details on the applied metrics.

(1)	*Ref:*	*Gesucht wurde auch im nahen Ausland.*
	Hypo:	Auch im nahen Ausland wurde gesucht.
		BLEU: **0.562** | WER: **0.667** | CER: **0.771**
(2)	*Ref:*	*Der Spatenstich fand im Oktober letzten Jahres statt.*
	Hypo:	Der Spatenstich fand letztes Jahr im Oktober statt.
		BLEU: **0.271** | WER: **0.500** | CER: **0.404**
(3)	*Ref:*	*Überlegungen die Lage in Zukunft zu verbessern sind in Planung.*
	Hypo:	Gedanken wie man die Lage zukünftig besser machen kann sind in Planung.
		BLEU: **0.159** | WER: **0.700** | CER: **0.532**

Thanks to recent advancements in paraphrasing systems based on neural machine translation (Thompson and Post, 2020b), the quality of paraphrases significantly increased. State-of-the-art paraphrasing systems not only cover synonym substitution but more advanced changes in sentence structures. Existing research improving translation metrics based on automatic paraphrasing (Bawden et al., 2020) focuses on the augmentation of refer-

ences and approaches to increase diversity to maximize coverage of the translation space.

Building upon the existing idea of augmenting the references for an improved BLEU metric by automatic paraphrasing, we extend this method to also generate paraphrases of the hypothesis. Our experiments show that this addition significantly improves the correlation of common metrics with human perception.

2 Related Work

The use of paraphrasing for evaluating machine translation systems to address some of the weaknesses of popular metrics has a long history. The metrics METEOR (Banerjee and Lavie, 2005), Meteor Universal (Denkowski and Lavie, 2014), and ParaEval (Zhou et al., 2006) support synonym matching, covering the simplest form of paraphrasing. A more recent approach ParaBLEU(Weston et al., 2021) includes a learned neural metric based on paraphrase representation learning, achieving state-of-the-art performance on the WMT Metrics Shared Task 2017 (WMT17) (Bojar et al., 2017). Bawden et al., 2020 showed that slight gains to the correlation with human judgment can be expected with automatic paraphrasing to generate additional references for BLEU. In a more general analysis from Freitag et al. 2020 the impact of well-chosen references on correlation with human judgment for English to German translation is analyzed. The researchers found that a precisely defined paraphrasing task executed by professional linguists increases the correlation compared to backtranslation or other automated methods.

3 Approach

Compared to existing approaches that introduce automatic paraphrasing to the calculation of the BLEU metric (Kanayama, 2003; Bawden et al., 2020), our approach is not limited to paraphrasing for the generation of diverse references but also generates paraphrases for the system output. Because the paraphrasing models based on machine translation typically have limited diversity of generated sentence structures, it is up to chance whether the sentence structure of a correct system output can be reproduced or not. With our approach of paraphrasing both the reference and the hypothesis ($\text{Para}_{\text{both}}$) we aim to increase the chances for an intended match and therefore limit the number of diverse paraphrases needed.

For the generation of paraphrases in German, we use a paraphrasing algorithm (Thompson and Post, 2020b) based on the Prism translation model (Thompson and Post, 2020a). This algorithm pushes the output away from the input in the lexical space by penalizing n-gram overlaps. The algorithm penalizes n-gram (1-, 2-, 3-, and 4-grams) overlaps by subtracting values from the output log probabilities of the NMT model before selecting candidates during beam search. With exponential weighting on the penalization, the method ensures penalizing larger n-gram overlaps more harshly than smaller ones. A parameter α controls how much the model pushes the output away from the input during decoding. We use the paraphrasing algorithm and the translation model with the parameters from Thompson and Post 2020b to sample the top n backtranslation candidates. This corresponds to $\alpha = 0.003$, $\beta = 4$, and a beam width of n. Examples of resulting paraphrases are reported in Appendix A.

With n additional generated paraphrases of the hypothesis and reference, we calculate $n+1$ metric values per sample for metrics supporting multiple references (e.g., BLEU). Single reference metrics (e.g., WER, CER) produce $(n+1)^2$ values. We explore different methods to aggregate the resulting values.

4 Evaluation

We consider two annotated data sources to measure the impact of our method if applied to established metrics. The two datasets Human Sentence Ratings (*GER-HSR-1K*) and Online Transcription Ratings (*GER-OTR-691*), specifically created for this work, are described in the following sections.

4.1 Human Sentence Ratings

We suggest a novel rating system aiming at capturing the essential components for comparing reference sentences and corresponding hypotheses for Standard German. The rating system consists of three binary values and one discrete rating with values between 0 and 3:

- Hypothesis Grammar (case-insensitive and without punctuation)
- Hypothesis Punctuation
- Hypothesis Capitalization
- Semantic similarity rating

Table 2: Semantic similarity rating range.

Rating	Description
3	Reference message is completely and unambiguously captured in the hypothesis (even possible if there are grammar or spelling mistakes)
2	Virtually matching semantic meaning with only lack or abundance of insignificant details, misspelling of named entities is also considered an insignificant detail
1	Majority of reference message is captured and only a small significant part is semantically not matching with the reference
0	Majority of reference message is not captured

The first three binary values only indicate correctness or an error regarding grammar, punctuation, and capitalization. The fourth rating corresponds to a semantic comparison between the reference and the hypothesis as defined in Table 2.

Based on this approach, we annotated a dataset with 1000 samples, denoted as *GER-HSR-1K*. The samples originate from the two datasets Swiss Parliaments Corpus (Plüss et al., 2021), and SDS-200 (Plüss et al., 2022), containing Swiss German audio with Standard German transcriptions. The audio samples from these datasets were transcribed using a speech-to-text system, finetuned for Swiss German and based on the fairseq S2T (Ott et al., 2019; Wang et al., 2020) model. We replicate the Transformer baseline model architecture and training procedure from Plüss et al., 2022. With this model, we generate realistic samples occurring in Swiss German to German speech translation. However, the references and corresponding hypotheses were only considered if they have a Levenshtein text distance greater than zero in order to exclude identical reference/hypothesis pairs.

In Figure 1, we report the distribution of the semantic similarity ratings. Because the sentences are sampled from a speech-to-text model achieving state-of-the-art results for Swiss German, lower ratings are less frequent and almost half of the samples are rated as perfectly matching the semantic meaning of the reference.

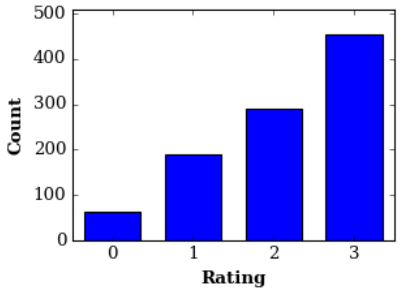

Figure 1: The distribution of the human-annotated semantic similarity rating in the *GER-HSR-1K* dataset.

Our human annotators found a grammar, punctuation, or capitalization error in only 3.2% of the *GER-HSR-1K* hypotheses. The models mostly produce grammatically sound sentences. For the analysis in this work, we only make use of the semantic similarity rating to evaluate our metric.

The fully annotated dataset is available online.[1]

4.2 Online Transcription Ratings

Through an online demo application showcasing a Swiss German speech-to-text system, we collected audio samples. After the transcription is shown to the user, they can voluntarily give feedback in the form of a discrete rating between 1 and 5 stars, with 1 star being the worst and 5 stars being the best rating. There is no further instruction given on how to rate the sentences. This dataset is extended by a human transcription to have a single reference ground-truth sentence.

Compared to the *GER-HSR-1K* dataset, the sentences collected in this dataset are on average shorter and less representative of naturally spoken language. Due to the lack of detailed instructions, the crowd-sourced ratings tend to be inconsistent. Additionally, because the ground-truth transcription is loosely based on the system hypothesis, little paraphrasing occurs.

A filtered version of this dataset with 691 samples is denoted as *GER-OTR-691*. This subset only contains pairs of references and hypotheses with a Levenshtein distance greater than zero.

4.3 Evaluation Method

To estimate and compare the quality of different metrics, we calculate correlations (linear: Pearson's r, monotonic: Kendall's τ) between metric results and human-annotated ratings. Specifically, we apply Kendall's Tau-like formulation defined

[1] https://www.cs.technik.fhnw.ch/i4ds-datasets

in the WMT18 Metrics Shared Task (Ma et al., 2018). The adaptation of Kendall's Tau coefficient is defined as follows:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (1)$$

Whether a comparison between human judgment and a metric of a pair of distinct samples, s_1 and s_2, is counted as concordant (Conc) or discordant (Disc) is defined in the following matrix, where $m(s_i)$ and $h(s_i)$ are the metric value and the human rating of the sample s_i, respectively:

		Metric		
		$m(s_1) < m(s_2)$	$m(s_1) = m(s_2)$	$m(s_1) > m(s_2)$
Human	$h(s_1) < h(s_2)$	Conc	Disc	Disc
	$h(s_1) = h(s_2)$	-	-	-
	$h(s_1) > h(s_2)$	Disc	Disc	Conc

This formulation means we exclude all human ties. In the case of non-identical human judgments, metric ties are always counted as $Discordant$. For this correlation coefficient to be consistent, we must ensure that human judgment and the metric have the same orientation, i.e., a higher score indicating higher transcription quality.

We use the sentence-level BLEU formula from sacreBLEU (Post, 2018) with exponential smoothing. For the WER and CER metrics, we employ the corresponding accuracy rate ($1 - \textit{Error-rate}$) to properly align the scores for Kendall's Tau-like formulation. We normalize the text by removing all punctuation and transforming all characters to lowercase.

5 Results

We analyzed different aggregation methods: maximum, minimum, average, or top-n-averaging. We found the minimum and the average to decrease the correlation with human judgment compared to the maximum. In some cases, averaging a subset of the top values outperformed the maximum, but the results were inconsistent. Because ties between metric values of two distinct samples are discouraged in the used Kendall's Tau-like formulation, the maximum, and an additional arbitrarily small random number added outperforms all other methods by decreasing the chances of ties to virtually zero. Statistically, in 50% of the cases where a tie would occur, we randomly hit the correct ranking. For a fair comparison, we report results without adding this random value. The added random number only helps to artificially increase Kendall's Tau-like values. It does not improve the metric in any useful way. All the results are reported with the maximum as the aggregation function.

Table 3: Kendall's τ and Pearson's r correlations on the *GER-HSR-1K* (HSR) and the *GER-OTR-691* (OTR) datasets. Para_{ref} (n) refers to the multi-reference version with n additional generated references. Metrics denoted as $\text{Para}_{\text{both}}$ (n) refer to our suggested extended version with the hypothesis and references augmented with n additional generated paraphrases.

Dataset	Metric	τ	r
HSR	WER	0.3472	0.4704
	Para_{ref} (11) WER	0.3725	0.4879
	$\text{Para}_{\text{both}}$ (6) WER	**0.5115**	**0.6137**
	CER	0.2632	0.3511
	Para_{ref} (11) CER	0.3081	0.3982
	$\text{Para}_{\text{both}}$ (6) CER	**0.4811**	**0.5513**
	BLEU	0.3167	0.3903
	Para_{ref} (11) BLEU	0.3798	0.4438
	$\text{Para}_{\text{both}}$ (6) BLEU	**0.4892**	**0.5872**
OTR	WER	0.5903	0.6133
	Para_{ref} (8) WER	0.5957	0.6191
	$\text{Para}_{\text{both}}$ (2) WER	**0.5972**	**0.6238**
	CER	0.6283	0.6516
	Para_{ref} (8) CER	0.6356	0.6590
	$\text{Para}_{\text{both}}$ (2) CER	**0.6369**	**0.6632**
	BLEU	0.5531	0.6419
	Para_{ref} (7) BLEU	0.5369	0.6270
	$\text{Para}_{\text{both}}$ (2) BLEU	**0.5578**	**0.6473**

The results for the two reference datasets are reported in Table 3. For both augmentation methods, we report results for the best number of paraphrases between 1 and 16. We show a significant improvement of the correlations for the *GER-HSR-1K* dataset if applying $\text{Para}_{\text{both}}$. With the augmentation limited to the references (Para_{ref}), we see a much lower improvement to the baseline. On the *GER-OTR-691* dataset we observe very limited gains. Because of the lack of paraphrasing occurring in this dataset, we did not expect a large improvement in the correlations.

In Figure 2, we show the impact of the number of paraphrases on the correlation with human judgment. Throughout all metrics six seems to be the best number of paraphrases for our method $\text{Para}_{\text{both}}$.

The parameterization of the underlying automatic paraphrasing model and the dataset impacts the number of paraphrases. Therefore, the chosen number of paraphrases should not be interpreted as a generally good pick. Additionally, we include the result of the WER metric with the augmentation limited to references (Para_{ref}). In this case, the strongest correlation between human ratings and automatic metrics is achieved at 11 paraphrases but is significantly lower than our parallel approach.

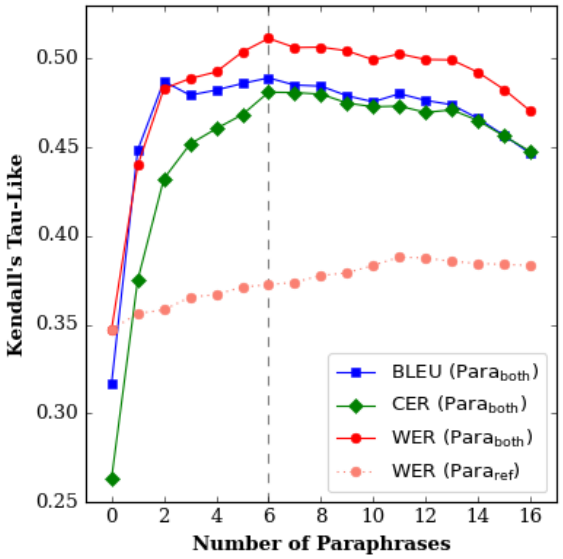

Figure 2: Results of the $\text{Para}_{\text{both}}$ method applied to three metrics on the *GER-HSR-1K* dataset with different numbers of paraphrases (for both the reference and the hypothesis). Additionally, the WER results with only the references augmented (Para_{ref}). Zero paraphrases correspond to no augmentation and the single-reference version of the metric.

In addition to the numeric correlation, we show a visual comparison of human ratings and metric values in Figure 3. Based on the single reference BLEU scores, we observe that the metric clearly underestimates a lot of samples with a high human rating of 2 or 3. The metric distributions of the samples with these high ratings are nearly indistinguishable. Our approach improves these distributions and better aligns the metrics to be linearly increasing with the human ratings. Because our approach combined with maximum aggregation can only increase the metric value, we overestimate some of the low-rated samples.

6 Conclusion

In this paper, we introduced $\text{Para}_{\text{both}}$, an augmentation method for translation metrics. We demonstrated a significant increase in correlation with human judgment for Swiss German to German speech translation. Based on state-of-the-art automatic paraphrasing, we produce different paraphrased versions of the reference and the hypothesis. Using our method, we improve the robustness of existing metrics by addressing paraphrasing that may arise in translation tasks.

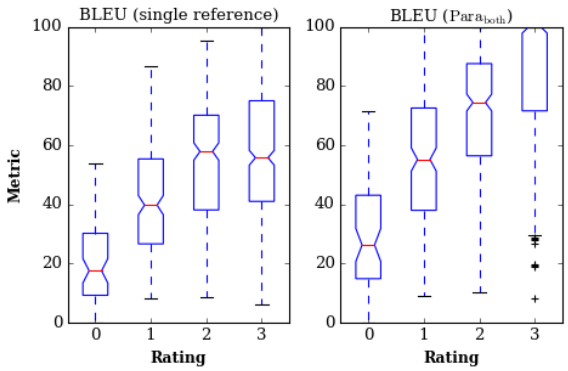

Figure 3: A comparison between the distributions of the single reference BLEU and the $\text{Para}_{\text{both}}(6)$ BLEU score for the different semantic similarity ratings in the *GER-HSR-1K* dataset.

Based on encouraging side experiments conducted on the WMT19 (Fonseca et al., 2019) dataset, we propose further investigation into the overall effectiveness of this approach in neural machine translation. We observed a relative increase of 52% on Kendall's Tau-like score with $\text{Para}_{\text{both}}$.

For future work, we also recommend exploring methods to mitigate the overestimation of low-rated samples. Our current method does not include any measures to reduce the risk of overestimation. Additionally, exploring other paraphrasing methods and their parameter space may also lead to more suitable paraphrases.

The novel dataset *GER-HSR-1K* is made publicly available to help advance the development of more appropriate metrics for speech translation, especially Swiss German.

Acknowledgements

This work was supported by HASLER Foundation within the project *Quality metrics for Swiss German speech recognition* [22031].

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with im-

proved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Hiroshi Kanayama. 2003. Paraphrasing rules for automatic evaluation of translation into Japanese. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 88–93, Sapporo, Japan. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. In *Proceedings of the Swiss Text Analytics Conference 2021, Winterthur, Switzerland, June 14-16, 2021 (held online due to COVID19 pandemic)*, volume 2957 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*, Online. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

Jack Weston, Raphael Lenain, Udeepa Meepegama, and Emil Fristed. 2021. Generative pretraining for paraphrase evaluation.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics.

A Paraphrase Examples

Ref:	*Gesucht wurde auch im nahen Ausland.*
	Es wurde auch im nahen Ausland gesucht. Es wurde auch im benachbarten Ausland gesucht. Auch im benachbarten Ausland wurde gesucht. Auch im benachbarten Ausland suchte man. Es wurde auch im nahen Ausland verfolgt. Es wurde auch im benachbarten Ausland verfolgt.
Hypo:	*Auch im nahen Ausland wurde gesucht.*
	Ebenfalls im benachbarten Ausland wurde gesucht. Auch im benachbarten Ausland wurde gesucht. Ebenfalls im benachbarten Ausland suchte man nach. Ebenfalls im benachbarten Ausland suchte man. Auch im benachbarten Ausland suchte man. Ebenfalls im nahen Ausland wurde gesucht.

Ref:	*Der Spatenstich fand im Oktober letzten Jahres statt.*
	Der Spatenstich geschah im Oktober letzten Jahres. Der Spatenstich geschah im Oktober vergangenen Jahres. Der Spatenstich geschah im vergangenen Oktober. Der Spatenstich fand letztes Jahr im Oktober statt. Es geschah im Oktober letzten Jahres. Es geschah im Oktober vergangenen Jahres.
Hypo:	*Der Spatenstich fand letztes Jahr im Oktober statt.*
	Der Spatenstich fand im letzten Jahr im Oktober statt. Der Spatenstich fand im vergangenen Jahr im Oktober statt. Der Spatenstich fand im letzten Jahr im vergangenen Oktober statt. Der Spatenstich fand im Oktober des vergangenen Jahres statt. Der Spatenstich fand im letzten Jahr im Oktober ab. Es fand im letzten Jahr im Oktober statt.

Ref:	*Überlegungen die Lage in Zukunft zu verbessern sind in Planung.*
	Überlegungen zur Verbesserung der Zukunftssituation sind geplant. Überlegungen zur Verbesserung der Situation in der Zukunft sind geplant. Überlegungen zur Verbesserung der Situation in der Zukunft sind in Planung. Erwägungen zur Besserung der zukünftigen Lage gibt es. Erwägungen zur Besserung der zukünftigen Lage sind geplant. Überlegungen wie zukünftig die Dinge besser gestaltet werden können, sind in Planung.
Hypo:	*Gedanken wie man die Lage zukünftig besser machen kann sind in Planung.*
	Gedanken darüber, wie zukünftig die Dinge besser gestaltet werden können, sind in Planung. Gedanken darüber, wie zukünftig die Dinge besser gestaltet werden können, befinden sich in Planung. Ideen, wie sich die Situation in der Zukunft verbessern ließe, befinden sich in Planung. Ideen, wie sich die Situation in der Zukunft verbessern ließe, entwickeln sich. Jetzt geht es darum, darüber nachzudenken, wie man die Lage zukünftig besser machen kann. Überlegungen zur Verbesserung der Situation in der Zukunft sind im Gange.