Reference-less Measure of Faithfulness for Grammatical Error Correction

Leshem Choshen[1] and Omri Abend[2]
[1] School of Computer Science and Engineering, [2] Department of Cognitive Sciences
The Hebrew University of Jerusalem
leshem.choshen@mail.huji.ac.il, oabend@cs.huji.ac.il

Abstract

We propose USIM, a semantic measure for Grammatical Error Correction (GEC) that measures the semantic faithfulness of the output to the source, thereby complementing existing reference-less measures (RLMs) for measuring the output's grammaticality. USIM operates by comparing the semantic symbolic structure of the source and the correction, without relying on manually-curated references. Our experiments establish the validity of USIM, by showing that (1) semantic annotation can be consistently applied to ungrammatical text; (2) valid corrections obtain a high USIM similarity score to the source; and (3) invalid corrections obtain a lower score.[1]

1 Introduction

Evaluation in Monolingual Translation, and particularly in Grammatical Error Correction (GEC) is a challenging research field, much due to the difficulty in integrating different types of rewriting operations into a single measure, and the vast number of valid outputs (Tetreault and Chodorow, 2008; Madnani et al., 2011; Chodorow et al., 2012; Bryant and Ng, 2015). These difficulties have recently motivated a number of proposals for new, improved reference-based measures (RBMs) (Dahlmeier and Ng, 2012; Felice and Briscoe, 2015; Napoles et al., 2015).

Nevertheless, the size and heterogeneity of the space of valid outputs per sentence often prohibits obtaining a reference set that covers this space well, thereby limiting the applicability of RBMs (Bryant and Ng, 2015). To address this we propose a semantic RLM, USIM, that operates by measuring the graph distance between the semantic representations of the source and the output. Reliable RLMs are appealing both in not relying on references, which are costly to collect, and in avoiding the biases incurred by selecting references that necessarily cannot exhaust the vast space of valid corrections.

Our proposal complements the RLM proposed by Napoles et al. (2016), which uses grammatical error detection techniques to assess the grammaticality of the output, and the work of Asano et al. (2017), who advocate the use of RLMs for fluency, grammaticality and meaning preservation, but state that a meaning preservation measure for GEC is currently lacking. A similar decomposition of output quality to its adequacy (similar to faithfulness) and fluency (related to grammaticality), has been used in machine translation (MT) evaluation (e.g., Banchs et al., 2015).

As a test case, we use the UCCA semantic scheme (Abend and Rappoport, 2013), motivated by its recent use in semantic evaluation of MT (Birch et al., 2016) and text simplification (Sulem et al., 2018) systems. Nevertheless, USIM can be easily adapted to other semantic schemes, such as AMR (Banarescu et al., 2013). USIM is conceptually related to RLMs developed for MT (Reeder, 2006; Albrecht and Hwa, 2007; Specia et al., 2009, 2010). Notably, XMEANT (Lo et al., 2014) compares the source to the output in terms of their semantic role labeling structures. Our use of UCCA is motivated by its wider coverage of predicate types, as opposed to MEANT's focus on verbal predicates, and UCCA's preservation of structure across translations (Sulem et al., 2015). See (Birch et al., 2016) for further discussion.

We conduct experiments to confirm USIM's validity. Specifically, we show that (1) UCCA can be consistently and automatically applied to learner language (LL) (§4.2), (2) USIM is not prone to unduly penalize valid corrections (§4.2), and (3) USIM assigns a lower score to corrections of poor

---

quality (§4.5). Our experiments also indicate that UCCA parsing technology is already sufficiently mature for an automatic variant of USIM to provide reliable results (§4.3).

2 Background

LL Annotation. While most linguistic theories propose that each learner makes consistent use of syntax (Huebner, 1985; Tarone, 1983), this use may not conform to the syntax of the learned language, or of any other known language. This entails difficulties in defining syntactic annotation for LL, as the annotated syntax differs between learners.

Syntactic schemes for LL annotate syntactically erroneous sentences in different ways. Berzak et al. (2016) and Ragheb and Dickinson (2012) annotate according to the syntax used by the learner, even if this use is not grammatical. Such annotation may be unreliable for measuring faithfulness, as GEC systems aim to alter these erroneous syntactic structures. Nagata and Sakaguchi (2016) take the opposite approach, and remain faithful to the syntax intended by the learner. This has also been the tradition in works on parser robustness (Bigert et al., 2005; Foster, 2004). However, such approach is prone to inconsistencies due to the variety of different syntactic structures that can be used to express a similar meaning.

In this paper, we use semantic annotation to structurally represent LL. Semantic structures are faithful to the intended meaning, and not to the formal realization, and thus face fewer conflicts where the syntactic structure used diverges from the one intended. We are not aware of any previous attempts to semantically annotate LL text.

The UCCA Scheme. UCCA is a semantic annotation scheme that builds on typological and cognitive linguistic theories. The scheme's aims are to provide a coarse-grained, cross-linguistically applicable representation. Importantly, UCCA's categories directly reflect semantic, rather than distributional distinctions. For instance, UCCA is not sensitive to POS distinctions: a Scene's main relation can be a verb but also an adjective ("He is **thin**") or a noun ("John's **decision**"). Indeed, Sulem et al. (2015) have found that UCCA structures are preserved remarkably well across English-French translations.

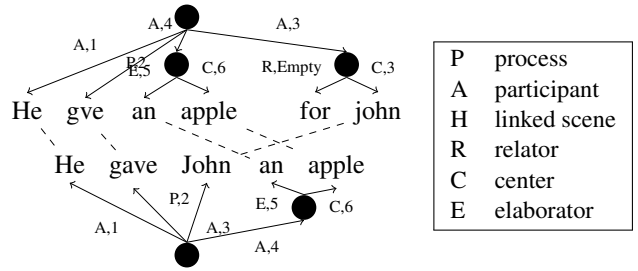UCCA structures are directed acyclic graphs, where the words correspond to (a subset of) their

Figure 1: UCCA structures of a learner language (top) and correction (bottom) including word alignments (dashed). On the edges are labels and numbers aligned to (top) or indexes (bottom). Precision is $\frac{7}{9}$ Recall is $\frac{7}{7}$.

leaves. The nodes of the graphs, called *units*, are either leaves or several elements jointly viewed as a single entity according to some semantic or cognitive consideration. The edges bear one or more categories, indicating the role of the sub-unit in the relation that the parent represents.

UCCA views the text as a collection of *Scenes* and relations between them. A Scene describes a movement, an action or a state which is persistent in time. Every Scene contains one main relation, zero or more *Participants*, interpreted in a broad sense to include locations, destinations and complement clauses, and *Adverbials*, such as manner or aspectual modifiers.

3 Semantic Faithfulness Measures

We start by defining a simplified measure, used for inter-annotator agreement (IAA). The measure compares two UCCA annotations over the same set of tokens. We then proceed to define USIM, which compares two UCCA structures over alignable but different sets of tokens.

IAA Measure. We define a similarity measure over UCCA annotations G_1 and G_2 that share their set of leaves (tokens) W. For a node v in G_1 or G_2, define its yield $yield(v) \subseteq W$ as its set of leaf descendants. Define a pair of edges $(v_1, u_1) \in G_1$ and $(v_2, u_2) \in G_2$ to be matching if $yield(u_1) = yield(u_2)$ and they have the same label. Labeled precision and recall are defined by dividing the number of matching edges in G_1 and G_2 by $|E_1|$ and $|E_2|$ respectively. *DAG F-score* is their harmonic mean. The measure collapses to the common parsing F-score if G_1, G_2 are trees.

The USIM Measure. Computing a faithfulness measure is slightly more involved, as the source sentence graph G_s and its correction G_c do not share the same set of leaves. We assume a (pos-

sibly partial, possibly many-to-1) alignment between G_s and G_c, $A \subset V_s \times V_c$.

An edge $(v_1, v_2) \in E_c$ is said to match an edge $(u_1, u_2) \in E_s$ if they have the same label and $(v_2, u_2) \in A$. Recall (Precision) is defined as the ratio of edges in E_s (E_c) that have a match in E_c (E_s) respectively, and F-score is their harmonic mean. We note that this measure collapses to the *DAG F-score* if A includes all pairs of nodes in E_s and E_c that have the same yield. See Figure 1.

In order to define the alignment between V_s and V_c, we begin by aligning the leaves (tokens) in V_s and V_c. Alignment is cast as a weighted bipartite graph matching problem. Edge weights are assigned to be the edit distances between the tokens. We note that aligning words in GEC (and other monolingual translation tasks) is much simpler than in MT, as most of the words are unchanged, deleted fully, added, or changed slightly. Denote the resulting leaf alignment with $A_l \subset Leaves_s \times Leaves_c$. We extend A_l to define the node alignment A, aligning each non-leaf $v \in V_s$ to the node $u \in V_c$ that maximizes

$$w(v, u) = \frac{|A_l \cap (yield(u) \times yield(v))|}{|yield(u)|}.$$

We exclude from A zero-weighted pairs. USIM is defined to be the F-score resulting from A. As the alignment may differ when aligning nodes from V_c to V_s and the other way around, we report USIM in both directions.

USIM is somewhat more relaxed than *DAG F-score*, as, unlike DAG F-score, it also aligns nodes whose yields are not in perfect alignment with one another. This relaxation is necessary, given that corrections often add or remove nodes, thus eliminating the possibility of a perfect alignment. In order to obtain comparable IAA scores, we report IAA using USIM as well.

For completeness, we replicate the protocol used by Sulem et al. (2015) for comparing the UCCA annotations of standard English-French translations, which we call Distributional Similarity (DISTSIM). For a given UCCA label l, $c_i(l)$ is the number of l-labeled UCCA edges in the i-th source sentence, and $d_i(l)$ is the number of l-labeled UCCA edges in its corresponding correction. We define DISTSIM(l) between these sentences to be $\frac{1}{N} \sum_{i=1}^{N} |c_i(l) - d_i(l)|$, where N is the total number of sentence pairs.

4 Experiments

We conduct four types of experiments to validate USIM, showing that: (1) semantic annotation can be consistently applied to LL through inter-annotator agreement (IAA) experiments; (2) a valid corrector scores high on USIM; (3) an automatic UCCA parser can reliably replace human annotation for USIM; (4) USIM is sensitive to changes in meaning.

4.1 Experimental Setup.

We train two UCCA annotators, the first author and a paid in-house annotator by annotating both LL and standard English passages, until a high enough agreement is reached (6 training hours). Training passages are excluded from the evaluation. We use UCCA's annotation guidelines[2] without any adaptations.

We experiment on 7 essays and their corrections, each comprising about 500 tokens (see supplementary material 1). In order to measure IAA, we assigned 4 of these essays to both annotators. In order to measure the faithfulness score for a valid correction, we annotate both the source and the manually corrected versions of 6 essays, 3 of which were annotated by both annotators.

4.2 The Faithfulness of Valid Corrections.

We obtain an IAA *DAG F-score* of 0.845 (Precision 0.834, Recall 0.857), which is comparable to the IAA reported for English Wikipedia texts by Abend and Rappoport (2013). As another point of comparison, we doubly annotate 3 corrected NUCLE (Dahlmeier et al., 2013) passages, obtaining a similar IAA. These results suggest that UCCA annotating LL does not degrade IAA: it can be applied as consistently to LL as to standard English.

Table 1 (left-hand side) presents the USIM scores obtained by comparing the NUCLE references and the source, or equivalently the score of a valid correction. To control for differences between the annotators, we explore both a setting where both sides are annotated by the same annotator, and a setting where they are annotated by different ones. As an upper bound on the score of a valid corrector (using different annotators), we also report the USIM IAA on source sentences.

Our results indicate that a valid correction obtains a score comparable to the IAA, which indi-

	USIM			DISTSIM	
	s→r	r→s	Avg	A+D	Scene
Different	0.85	0.83	0.84	0.96	0.93
Same	0.92	0.91	0.92	0.97	0.96
IAA	0.85	0.81	0.83	-	-
SAR15	-	-	-	0.95	0.96

Table 1: The faithfulness of valid corrections. The left-hand side presents USIM, where s→r is the setting where alignment is computed from the source to the reference, r→s is the other way around, and Avg is their average. The right-hand side presents DISTSIM for the UCCA categories Participants and Adverbials together (A+D), and for Scenes (Scene). Rows indicate whether the same annotator annotated the source and reference or not. For comparison, the IAA row is the IAA computed using USIM. Results show that the valid corrector's faithfulness is comparable with IAA. SAR15 are reported by Sulem et al. on English-French translations; similarity is comparable to ours.

cates that USIM is indeed insensitive to the surface divergence between a source sentence and its valid corrections. Finally, we compute the DISTSIM measure between the source and reference sentences (Table 1, right-hand side), obtaining similar results to those obtained by Sulem et al. (2015). It suggests that on a coarse grained level, UCCA structures are as robust to grammatical error corrections as they are to translation from English to French, which was shown to be very robust, specifically more robust than syntactic representation (Sulem et al., 2015).

4.3 Automatic USIM.

We experiment with an automatic variant of USIM, where UCCA structures are parsed automatically. We use the TUPA parser (Hershcovich et al., 2017) to generate UCCA structures, instead of the human annotators. Otherwise the setup is as above. TUPA is used with its biLSTM model, trained on the UCCA English Wikipedia corpus.

We obtain a USIM score of 0.7 between the parses of the reference correction and the source, which is comparable to the parser's reported performance (0.73 in-domain, 0.68 out-of-domain), despite not performing any domain adaptation to LL. That is, the UCCA parses of the source and the correction are roughly as similar to each other as they are to their gold standard parse. This supports the hypothesis that semantic parsing technology is sufficiently mature to be applicable to USIM. Results also suggest an improvement in parsing performance may further improve these scores.

4.4 Sensitivity to Error Types

To provide another perspective on automated USIM's behaviour, we examined the measure's sensitivity to different error types, using a methodology from an ongoing work. For each NUCLE sentence and set of edits (replacements of substrings that contain an error by corrected ones. Such edit for the example in fig. 1 might be "gva" → "gave", with type *spelling*), we sample an order in which edits are applied. We select the source randomly to be one of the resulting sentences. We then compare the difference in USIM before and after applying each edit, and average these differences by the applied edit type. We denote the average difference in USIM due to correction of errors of type t with Δ_t. The hypothesis is that Δ_t should be close to 0 for all t, as edits are manual and are thus assumed to be faithful. We focus on edit types with high $|\Delta_t|$ to better understand where USIM fails. See table 2 in the supplementary material for complete results.

We find that among the 5 most penalized error types by USIM are "unclear meaning" and corrections of type "other", that fit no specific type; these corrections may actually change the meaning of the original sentence. In the most penalized and most rewarded changes we see "Dangling Modifier", "Pronoun Reference" and "Word Tone" errors, the first usually changes a word into a more complex structure and the latter two the opposite. Such changes alter the lower levels of the UCCA structure (near the leaves); a similarity measure that focuses on the top of the DAG, or one that performs a better lexical semantic abstraction, may address this sensitivity. Corrections of incorrect word order are also highly rewarded (high Δ_t), probably due to parser performance (the UCCA structures themselves are not affected by word order). Training the parser with LL annotated data may address this sensitivity.

Among the most rewarded changes we also see errors of replacing rare or misconstructed words with proper English words (Acronym and Mechanical errors). We assume this is due to parser performance, as TUPA only extracts features over complete words, and has no character-level encoding at this point. Thus, all misconstructed words fall into an out-of-vocabulary category and can only be labeled by the context.

Lastly, adding a missing verb is shown to be highly rewarded. Under the UCCA guidelines, a

missing verb should be annotated as an implicit unit, but as TUPA does not generate implicit units, it is not surprising that when corrections transforms an implicit unit into an explicit word, the parser's output changes (and hence USIM). Future improvements to TUPA may address this.

4.5 Sensitivity to Unfaithfulness.

We have shown that UCCA is insensitive to differences between a source sentence and its valid correction. We now present an evaluation of the sensitivity of USIM to proposed corrections that diverge semantically from the source. A semantic measure is, by its definition, sensitive to variation in the semantic dimensions which it encodes. In UCCA's case, these distinctions focus on predicate-argument structures, the inter-relations between them, and the semantic heads of complex arguments. These distinctions are widely regarded as fundamental in the NLP and linguistic literature.

In order to empirically validate this claim, we present an experiment which shows that corrections of a fairly low quality indeed receive a much lower USIM faithfulness score. Current state-of-the-art systems rarely alter the source sentences enough to yield semantically unfaithful outputs.[3] Consequently, their human rankings are not determined by their semantic faithfulness, rendering them unuseful for validating USIM. We instead experiment with 5 partially trained correctors, trained and evaluated on the JFLEG corpus (Napoles et al., 2017) by Sakaguchi et al. (2017).

USIM is computed automatically for each system's output on 754 source sentences. Low faithfulness results are expected, as these outputs include major changes, sometimes deleting full phrases from the output or changing every other word. Indeed, automatic USIM obtains scores of 0.32-0.39 for 4 of the systems, and 0.19 for the system that obtains the lowest GLEU (Napoles et al., 2015) score. For completeness, we run USIM on the 4 references provided by JFLEG for each source and obtain scores of 0.72-0.78, suggesting the domain change is not the reason for the low USIM score.

Taken together, these results indicate that USIM, even in its automatic variant, is sensitive to semantic changes. Consider the example:

USIM assigns the reference 0.71 and only 0.33

Source	the good student must know how to understand and work hard to get the iede.
Reference	A good student must be able to understand and work hard to get the idea.
Corrector	The good student must know how to understand and work hard to get on.

to the corrector. Moreover, although the reference makes more word changes than the proposed correction, it still obtains a higher USIM score.

5 Conclusion

We propose a measure of semantic faithfulness of a correction to the source, thereby avoiding the pitfalls of reference-based evaluation. We believe that using RLMs in conjunction with RBMs in the training and development of GEC systems will better address the challenge of over-conservatism, and the high costs of acquiring many references.

Future work will conduct user studies to assess the relative importance of different evaluation criteria. Specifically, we will explore to what extent users are tolerant to invalid changes to the sentence's structure, i.e., violation of conservatism, relative to their tolerance to invalid changes to the sentence's meaning, i.e., violation of faithfulness. A better understanding of how these interact may lead to improved semantic evaluation that will alleviate the need for a high number of references.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *ACL (1)*, pages 228–238.

Joshua Albrecht and Rebecca Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proc. of ACL*, volume 45, pages 880–887.

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced

---

[3]We demonstrate this in a paper currently under review.

with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 343–348.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. Association for Computational Linguistics.

Johnny Bigert, Jonas Sjöbergh, Ola Knutsson, and Magnus Sahlgren. 2005. Unsupervised evaluation of parser robustness. In *CICLing*, pages 142–154. Springer.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. Hume: Human ucca-based evaluation of machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *ACL (1)*, pages 697–707.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel R Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *COLING*, pages 611–628. Citeseer.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *HLT-NAACL*, pages 578–587.

Jennifer Foster. 2004. Parsing ungrammatical input: an evaluation procedure. In *LREC*. Citeseer.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for ucca. *arXiv preprint arXiv:1704.00552*.

Thorn Huebner. 1985. System and variability in interlanguage syntax. *Language Learning*, 35(2):141–163.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. 2014. Xmeant: Better semantic mt evaluation without reference translations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 765–771, Baltimore, Maryland. Association for Computational Linguistics.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 508–513. Association for Computational Linguistics.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase structure annotation and parsing for learner english. In *Proc. of ACL*, pages 1837–1847.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*.

Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *COLING (Posters)*, pages 965–974.

Florence Reeder. 2006. Measuring mt adequacy using latent semantic analysis. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts*, pages 176–184.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Grammatical error correction with neural reinforcement learning. *arXiv preprint arXiv:1707.00299*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Elior Sulem, Omri Abend, and Ari Rappoport. 2015. Conceptual annotations preserve structure across translations: A french-english case study. *Proceedings of S2MT 2015*, page 11.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural annotation for text simplification. In *Proc. of NAACL*. To appear.

Elaine Tarone. 1983. On the variability of interlanguage systems. *Applied linguistics*, 4(2):142–164.

Joel R Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 24–32. Association for Computational Linguistics.