What's in a Domain? Analyzing Genre and Topic Differences in Statistical Machine Translation

Marlies van der Wees **Arianna Bisazza** **Wouter Weerkamp**[◊] **Christof Monz**
Informatics Institute
University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl
[◊]904Labs, Amsterdam
wouter@904labs.com

Abstract

Domain adaptation is an active field of research in statistical machine translation (SMT), but so far most work has ignored the distinction between the *topic* and *genre* of documents. In this paper we quantify and disentangle the impact of genre and topic differences on translation quality by introducing a new data set that has controlled topic and genre distributions. In addition, we perform a detailed analysis showing that differences across topics only explain to a limited degree translation performance differences across genres, and that genre-specific errors are more attributable to model coverage than to suboptimal scoring of translation candidates.

1 Introduction

Training corpora for statistical machine translation (SMT) are typically collected from a wide variety of sources and therefore have varying textual characteristics such as writing style and vocabulary. The test set, on the other hand, is much smaller and usually more homogeneous. The resulting mismatch between the test data and the majority of the training data can lead to suboptimal translation performance. In such situations, it is beneficial to adapt the translation system to the translation task at hand, which is exactly the challenge of domain adaptation in SMT.

The concept of a *domain*, however, is not unambiguously defined across existing domain adaptation methods. Commonly used interpretations of domains neglect the fact that *topic* and *genre* are two distinct properties of text (Lee and Myaeng, 2002; Stein and Meyer Zu Eissen, 2006). Two texts can discuss a similar topic, but using different styles. Since most work on domain adaptation in SMT uses in-domain and out-of-domain data that differ on both the topic and the genre level, it is unclear whether the proposed solutions address topic or genre differences.

In this work we take a step back and disentangle the concepts topic and genre, then we analyze and quantify their effect on SMT, which we believe is a necessary step towards further improving domain adaptation for SMT. Concretely, we address the following questions:

(i) Can we clarify the ambiguous use of the concept *domain* with regard to adaptation in SMT?

(ii) Which of two intrinsic text properties, topic and genre, presents a larger challenge to SMT?

(iii) To what extent do topic and genre differ with respect to SMT model coverage and observed out-of-vocabulary (OOV) types?

To answer these questions, we introduce a new data set with controlled topic-genre distributions, which we use for an in-depth analysis of the impact of topic and genre differences on SMT.

2 Topic and genre differences in SMT

The definition of a domain varies across work on domain adaptation and is often imprecise. In this work we avoid using this ambiguous term, and instead focus on the text properties topic and genre.

Topic is the general subject of a document. Topics can be determined on multiple levels, ranging from very broad to more detailed. Examples of topics include sports, politics, and science (high-level), or football and tennis (low-level).

Genre is harder to define, as there is no single definition in literature (Swales, 1990; Karlgren,

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 560–566,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

Topic	Newswire sentence	User-generated sentence
Culture	The 12 contestants competed during a May 3rd Prime before a panel of judges and millions of viewers across the Arab world.	Your program's name is "Arab Idol", which is in English, and you allowed Barwas to participate and represent Iraq while she sings in Kurdish!!!
Economy	Yemen is mulling the establishment of 13 industrial zones across its six planned administrative regions in a bid to stimulate development and create job opportunities.	What development in Yemen are you talking about? We will continue to call for freedom until independence and liberation and the routing of the northern occupation from our lands.

Table 1: English-side samples from the Gen&Topic data set. All pairs of newswire (NW) and user-generated (UG) fragments in the data set discuss the same article and are topically related.

> 2004). Based on previous definitions, Santini (2004) concludes that the term genre is used as a concept complementary to topic, covering the non-topical text properties function, style, and text type. Like topics, genres can also exhibit different levels of granularity (Lee, 2001). Examples of genres include formal or informal text (high-level), and newswire, editorials, and user-generated text (low-level).

Topic and genre are both intrinsic properties of texts, but most work on domain adaptation uses provenance or subcorpus information to adapt SMT systems to a specific translation task (Foster and Kuhn, 2007; Duh et al., 2010; Bisazza et al., 2011; Sennrich, 2012; Bisazza and Federico, 2012; Haddow and Koehn, 2012, among others). In recent years, some work has explicitly addressed topic adaptation for SMT (Eidelman et al., 2012; Hewavitharana et al., 2013; Hasler et al., 2014a; Hasler et al., 2014c) using latent Dirichlet allocation (Blei et al., 2003). While Hasler et al. (2014b) showed that provenance and topic can serve as complements to each other, the effects of genre and topic on SMT have not been systematically studied.

3 The Gen&Topic benchmark set

To analyze the impact of genre and topic differences in SMT, we need a test set where both dimensions are controlled as much as possible. Unfortunately, currently available and commonly used benchmarks meet this requirement only to a limited degree. For instance, while the NIST OpenMT sets do contain documents drawn from two genres, newswire and web, both genres exhibit a different distribution over topics, i.e., the same topic might not be equally represented across genres, and vice versa.

To overcome this limitation, we introduce a new Arabic-English parallel benchmark set, the Gen&Topic data set, that contains documents with controlled topic and genre distributions. This benchmark set consists of manually translated news articles crawled from the web with their corresponding, manually translated readers' comments and thus comprises the genres *newswire* (NW) and *user-generated* (UG) text. Since each pair of NW and UG documents originates from the same article, we can assume that both documents discuss the same topic, for which labels are provided by the source websites. By including comparable numbers of tokens per genre for each article, we enforce equal topic distributions across the genres. Two examples of NW-UG pairs are shown in Table 1. Note that the selected UG sentences in the Gen&Topic data set are well-formulated comments rather than dialog-oriented content such as SMS or chat messages, which pose substantially larger challenges to SMT than the Gen&Topic comments (van der Wees et al., 2015).

		Genre		
Topic		NW	UG	Total
Culture	segments tokens	654 15.5K	507 14.9K	1161 30.4K
Economy	segments tokens	500 16.0K	578 15.5K	1078 31.5K
Health	segments tokens	384 9.7K	319 9.3K	703 19.1K
Politics	segments tokens	494 15.8K	646 15.8K	1140 31.6K
Security	segments tokens	532 16.1K	826 15.9K	1358 32.0K
Total	segments tokens	2564 73.2K	2876 71.3K	5440 144.5K

Table 2: Statistics of the Arabic-English Gen&Topic data set containing five topics and two genres: newswire (NW) and user-generated (UG) text. Tokens are counted on the Arabic side.

For parameter estimation purposes, we split the complete benchmark into a development and a test set, such that the development set contains approximately one-third of the data, while ensuring that articles in each set originate from non-overlapping time periods. Table 2 lists the specifications of the complete benchmark, which we make available for download[1].

4 Quantifying the impact of genre and topic differences on SMT

To quantify the impact of multiple genres and topics in a test corpus, we run a series of experiments in which we measure translation quality, model coverage, and observed OOV types.

4.1 Translation quality

We first run a translation experiment on the Gen&Topic test set using our in-house phrase-based SMT system similar to Moses (Koehn et al., 2007). Features include lexicalized reordering, linear distortion with limit 5, and lexical weighting. In addition, we use a 5-gram linearly interpolated language model, trained on 1.6B words with Kneser-Ney smoothing (Chen and Goodman, 1999), that covers all topics and genres contained in the benchmark. We tune our system on the Gen&Topic development set using pairwise ranking optimization (PRO) (Hopkins and May, 2011).

Naturally, performance differences across topics and genres depend on the degree to which both are represented in the parallel training data. To allow for fair comparison, we down-sample our available training data to be as balanced as possible in terms of topics and genres. The resulting system is trained on approximately 200K sentence pairs with 6M source tokens per genre, as much as is available for UG. All data originates from the same web sources as the documents in the benchmark. Our more competitive system (van der Wees et al., 2015) that uses also LDC-distributed data yields slightly higher BLEU scores, but is more favorable for NW than for UG translation tasks. Due to the strict data requirements in terms of topic and genre distributions, as well as the availability of sizable parallel training data, our current experimental set-up covers Arabic-English only.

Table 3 compares BLEU scores (Papineni et al., 2002, 1 reference) of the Gen&Topic data, split down by topics and genres. We observe that translation performance fluctuates much more across genres than across topics: There is a large gap of 3.9 BLEU points between NW and UG, which can be entirely attributed to actual genre differences given the construction of the Gen&Topic data set and the use of down-sampled training data. On the other hand, the gap between different topics is only 0.6 BLEU points on average, and at most 1.1 (between culture and politics). A translation quality gap between genres has also been observed in past OpenMT evaluation campaigns. However, as the NIST benchmarks have not been controlled for topics across genres, it is unclear to what extent this gap can be attributed to genre differences.

	NW	UG	All	
Culture	19.2	17.6	19.3	
Economy	19.9	15.9	18.9	
Health	19.3	17.7	18.8	Avg. diff.: ±0.6
Politics	21.3	13.6	18.2	
Security	19.3	16.2	18.5	
All	19.9	16.0	18.9	
	Avg. diff.: ±3.9			

Table 3: Arabic-to-English BLEU scores on the Gen&Topic test set (1 reference translation) per topic-genre combination. Tuning was done on the complete Gen&Topic development set. Variations in translation quality are represented by average pairwise BLEU score differences.

4.2 Model coverage analysis

Next, to explain the large performance gap between genres, we analyze the phrase lengths within Viterbi translations, source phrase and phrase pair recall, and phrase pair OOV of the Gen&Topic test set (Table 4).

Average source-side phrase length We first compute the average number of source words contained in the phrases that our SMT system uses to produce the 1-best translations for the Gen&Topic test set. One can see that UG is translated with shorter phrases than NW, and that differences between genres are more pronounced than among topics. This difference, in turn, can be due to unreliable translation probabilities but also to the mere lack of translation options in the models. We quantify the impact of the latter by measuring phrase recall on each test portion.

Phrase recall and phrase pair OOV To compute phrase recall, we first automatically word-

[1] http://ilps.science.uva.nl/resources/gen-topic/

Gen&Topic portion	BLEU	Avg.phr. length	Source phrase recall				Src-trg phrase pair recall				Phr.pair OOV
			1	2	3	4+	1	2	3	4+	
NW	19.9	1.45	99.3	81.4	41.8	7.1	73.8	39.4	13.7	1.8	71.5
UG	16.0	1.38	97.2	74.7	36.0	6.3	56.2	28.8	8.7	1.1	76.0
Culture	19.3	1.39	98.2	77.6	36.5	5.3	66.2	35.2	10.7	1.2	74.2
Economy	18.9	1.42	98.4	78.7	39.4	6.5	65.3	33.5	10.9	1.4	73.8
Health	18.8	1.41	98.3	76.6	37.1	5.4	64.5	33.5	11.0	1.2	75.2
Politics	18.2	1.41	98.1	78.6	39.8	7.7	60.8	33.1	11.2	1.5	73.4
Security	18.4	1.42	97.6	77.0	40.2	8.4	62.7	33.3	11.6	1.8	73.3

Table 4: Impact of genre and topic differences on various indicators of SMT model quality.

align the test set and extract from it a set of reference phrase pairs using the same procedure applied to the training data. Then, we count the number of reference phrase pairs whose source side is covered by the translation models (*source phrase recall*) and the number of reference phrase pairs that are fully covered by the translation models (*source-target phrase pair recall*). Formally, we define the set of source-matching phrases as:

$$M^S = \{(\bar{f}, \bar{e}) \mid (\bar{f}, \bullet) \in P_{test} \wedge (\bar{f}, \bullet) \in P_{train}\},$$

where P_d refers to the set of phrase pairs $(\bar{f}, \bar{e})$ that can be extracted from corpus d. Source phrase recall R_n^S for phrases of length n is then defined as:

$$R_n^S = \frac{\sum_{(\bar{f},\bar{e}) \in M^S \wedge |\bar{f}|=n} c_{test}(\bar{f}, \bar{e})}{\sum_{(\bar{f},\bar{e}) \in P_{test} \wedge |\bar{f}|=n} c_{test}(\bar{f}, \bar{e})}, \quad (1)$$

where $c_{test}(\bar{f}, \bar{e})$ denotes the frequency of phrase pair $(\bar{f}, \bar{e})$ in the test set. Analogously, we define the set of source-target-matching phrase pairs as:

$$M^{S,T} = \{(\bar{f}, \bar{e}) \mid (\bar{f}, \bar{e}) \in P_{test} \wedge (\bar{f}, \bar{e}) \in P_{train}\}$$

and the source-target phrase pair recall $R_n^{S,T}$ for phrases of length n as:

$$R_n^{S,T} = \frac{\sum_{(\bar{f},\bar{e}) \in M^{S,T} \wedge |\bar{f}|=n} c_{test}(\bar{f}, \bar{e})}{\sum_{(\bar{f},\bar{e}) \in P_{test} \wedge |\bar{f}|=n} c_{test}(\bar{f}, \bar{e})}. \quad (2)$$

Finally, we call *phrase pair OOV* the portion of reference phrase pairs that are not covered by the translation models, that is: $1 - \sum_n^N R_n^{S,T}$, where N is the phrase limit used for phrase extraction.

The results of our analysis, broken down by source phrase length, show that source phrase recall is much lower in UG than in NW, while variations among topics are only very small. The stronger impact of genre differences is even more visible on phrase pair recall: for instance, our system knows the correct translation of 73.8% of the single-source-word phrase pairs in the NW genre. In UG this is only 56.2%, despite the equal amounts of training data per genre in our system. These figures suggest that model coverage—both mono- and bilingual—is an important reason for the low SMT quality on UG data.

Most existing approaches to domain adaptation focus on domain-sensitive scoring or selection of existing translation candidates (Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011; Chen et al., 2013, among others). This strategy is supported by the error analysis of Irvine et al. (2013), who show that scoring errors are more common across domains than errors caused by OOVs, in the source as well as the target language. Across genres however, our results in Table 4 show that both word-level and phrase-level OOVs are a more likely explanation for the performance differences. This stresses the need to address model coverage, for example by paraphrasing (Callison-Burch et al., 2006) or translation synthesis (Irvine and Callison-Burch, 2014).

4.3 Manual OOV analysis

To get a better understanding of the OOVs observed for the genres and topics in the Gen&Topic set, we perform a fine-grained manual analysis[2]. For this analysis a bilingual speaker manually annotated 500 sentences on the source side (equally distributed over genres and topics) to identify the class of each OOV. Annotations are done for top and sub-level classes (e.g., replaced letter, which

[2] Available with the benchmark data at `http://ilps.science.uva.nl/resources/gen-topic/`

Arabic OOV	English translation	Explanation of OOV	Main OOV class
داعش	ISIL	New proper noun	Rare but correct (Rare)
هينسوا	(they) will forget	Dialectal future tense	Dialectal forms (Dial)
يقدسون	(they) revere	Third person plural present tense	Morphological variants (Morph)
توفيرالوظائف	creationofjobs	Missing blank	Spelling errors (Spell)
المتطوعيين	volunteeeers	Wrong but understandable spelling	Colloquialisms (Coll)

Table 5: Examples of OOVs observed in the Gen&Topic set with their respective main OOV class.

Gen&Topic	OOV type					
portion	Rare	Dial	Morph	Spel	Coll	Other
NW	77.8	0.0	16.7	5.6	0.0	0.0
UG	9.8	9.0	17.2	42.6	12.3	9.0
Culture	17.4	0.0	17.4	52.2	8.7	4.3
Economy	13.8	0.0	34.5	31.0	13.8	6.9
Health	15.8	10.5	15.8	36.8	10.5	10.5
Politics	25.0	25.0	12.5	25.0	0.0	12.5
Security	23.5	8.8	5.9	41.2	14.7	5.9

Table 6: Error percentages per Gen&Topic portion of main OOV classes, see Table 5 for explanation. Other events include words that are not understandable or occur in the phrase table but only captured in a different context.

is a subclass of spelling errors). In total, we consider 17 subclasses which we group into five main classes, see Table 5 for examples.

Table 6 shows the type level percentages[3] for each main OOV class per genre or topic. When comparing the two *genres*, a number of observations emerge. Firstly, rare but correct words (e.g., proper nouns and technical terms, both regular issues for adaptation in SMT) make up the vast majority of the OOVs in NW, but are relatively infrequent in UG. By contrast, OOVs containing unseen morphological variants are equally common in both genres. Although complex morphology is language-specific, a rare morphological word in Arabic often maps to a rare multi-word phrase in English, resulting in phrase-level OOVs. Next, not entirely surprising, the majority of OOVs in UG are due to spelling errors. Finally, OOVs assigned to the remaining classes are never observed in NW but occasionally occur in UG.

Next, a comparison of the main OOV classes among the various *topics* shows a few notable distributions. Dialectal forms, for example, are rare in all topics except politics, where they are commonly observed in the form of Egyptian future tense. This can be explained by the presence of news articles about elections in Egypt in the Gen&Topic set. Next, while spelling errors are common in all topics, its abundance is most prominent in culture. Most spelling errors concern missing or inserted blanks, suggesting that comments are likely written on mobile devices. Finally, unseen morphological variants are more frequent in economy than in other topics, however with no conclusive explanation.

5 Conclusions and implications

Despite the fact that domain adaptation is an active field of research in SMT, there is little consensus on what exactly constitutes a domain. By introducing and analyzing a new benchmark with balanced topic and genre distributions, we have shown that earlier findings explaining the differences across topics only explain to a limited degree translation performance differences across genres. Our analysis shows that genre-specific errors are more attributable to model-coverage errors than to suboptimal scoring of existing translation candidates. This suggests that future work on improving SMT across genres needs to investigate approaches that increase model coverage. Our fine-grained manual error analysis at the word level also suggests that source coverage could benefit from text normalization (Bertoldi et al., 2010). Finally, we make both our benchmark and the manual OOV annotations publicly available.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213. We thank Rachel Cotterill, Nigel Dewdney, and the anonymous reviewers for their valuable comments.

[3] We also collected token level frequencies which are very similar to the listed type level statistics, except for a small number of repeatedly occurring proper nouns.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2010. Statistical machine translation of texts with misspelled words. In *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419.

Arianna Bisazza and Marcello Federico. 2012. Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 439–448.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the 8th International Workshop on Spoken Language Translation*, pages 136–143.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1285–1293.

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT 2010)*, pages 243–250.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 115–119.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014a. Dynamic topic adaptation for phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of The Association for Computational Linguistics*, pages 328–337.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014b. Combining domain and topic adaptation for SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 139–151.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014c. Dynamic topic adaptation for SMT using distributional profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 445–456.

Sanjika Hewavitharana, Dennis Mehay, Sankaranarayanan Ananthakrishnan, and Prem Natarajan. 2013. Incremental topic-based translation model adaptation for conversational spoken language translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 697–701.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Stefan Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Jussi Karlgren. 2004. The wheres and whyfores for studying text genre computationally. In *Workshop on Style and Meaning in Language, Art, Music, and Design*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In

Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

David Y.W. Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717.

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.

Marina Santini. 2004. State-of-the-art on automatic genre identification. Technical Report ITRI-04-03, Information Technology Research Institute, University of Brighton.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549.

Benno Stein and Sven Meyer Zu Eissen. 2006. Distinguishing topic from genre. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, pages 449–456.

John M. Swales. 1990. *Genre Analysis*. Cambridge University Press., Cambridge, UK.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2015. Five shades of noise: analyzing machine translation errors in user-generated text. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*.