Neural Data-to-Text Generation Based on Small Datasets: Comparing the Added Value of Two Semi-Supervised Learning Approaches on Top of a Large Language Model

Chris van der Lee
Tilburg University
Tilburg Center for Cognition and Communication
c.vdrlee@uvt.nl

Thiago Castro Ferreira
Universidade Federal de Minas Gerais
Faculdade de Letras
thiagocf05@ufmg.br

Chris Emmery
Tilburg University
Department of Cognitive Science and Artificial Intelligence
c.d.emmery@uvt.nl

Travis J. Wiltshire
Tilburg University
Department of Cognitive Science and Artificial Intelligence
t.j.wiltshire@uvt.nl

Emiel Krahmer
Tilburg University
Tilburg Center for Cognition and Communication
e.j.krahmer@uvt.nl

Action Editor: Miguel Ballesteros. Submission received: 14 July 2022; revised version received: 28 February 2023; accepted for publication: 16 April 2023.

https://doi.org/10.1162/coli_a_00484

This study discusses the effect of semi-supervised learning in combination with pretrained language models for data-to-text generation. It is not known whether semi-supervised learning is still helpful when a large-scale language model is also supplemented. This study aims to answer this question by comparing a data-to-text system only supplemented with a language model, to two data-to-text systems that are additionally enriched by a data augmentation or a pseudo-labeling semi-supervised learning approach.

Results show that semi-supervised learning results in higher scores on diversity metrics. In terms of output quality, extending the training set of a data-to-text system with a language model using the pseudo-labeling approach did increase text quality scores, but the data augmentation approach yielded similar scores to the system without training set extension. These results indicate that semi-supervised learning approaches can bolster output quality and diversity, even when a language model is also present.

1. Introduction

Neural NLG methods are notoriously data hungry, and rely on large-scale datasets that typically require large amounts of effort and resources to construct (Gkatzia 2016). The fact that such datasets are rare and difficult to develop creates a so-called **data bottleneck** (Oraby et al. 2019). Due to the lack of large datasets, many neural NLG approaches rely on relatively small datasets, which not only affects output quality, but also output diversity (Holtzman et al. 2020).

One of the NLG subtasks that especially suffers from the consequences of small-scale datasets is data-to-text generation: The task of producing adequate, fluent and natural language text from non-linguistic structured data (Gatt and Krahmer 2018). (Supervised) neural data-to-text NLG involves the collection of parallel data–text datasets, aligning data and linguistic realizations of these data. However, collecting these datasets is difficult because sets of texts and corresponding data are not a common natural occurrence (Shimorina, Khasanova, and Gardent 2019). On the other hand, *unpaired* texts and data are significantly more common and easily collected (Qader, Portet, and Labbé 2019). While these unpaired texts and data do not lend themselves to supervised data-to-text generation, they can be utilized by means of **semi-supervised learning**, the process of training a model on existing data–text pairings, and having this model create more synthetic pairings for the training set (see also Figure 1).

Although more data generally leads to better performing machine learning systems, it is unclear to what extent a system benefits from including (possibly imperfect) synthetic data. Caution is advised when working with such data, as the inclusion of synthetic data may reinforce a model's mistakes (He et al. 2020). This issue becomes further exacerbated in the context of data-to-text generation, where neural models can be prone to issues such as hallucination, repetition, and omission of data (Faille, Gatt, and Gardent 2020). However, previous data-to-text generation studies suggest that semi-supervised learning increases output quality compared with a supervised approach, especially when the labeled dataset size is small (e.g., Qader, Portet, and Labbé 2019; Schmitt et al. 2020; Su, Huang, and Chen 2020; Tseng et al. 2020; Chang, Demberg, and Marin 2021). Besides improved output quality, adding more training examples using semi-supervised learning also might increase the language diversity of the output.

Recent developments in large-scale language models, utilizing the Transformers architecture (Vaswani et al. 2017), may offer an alternative solution to the data bottleneck. They leverage data from various domains to supplement the in-domain information

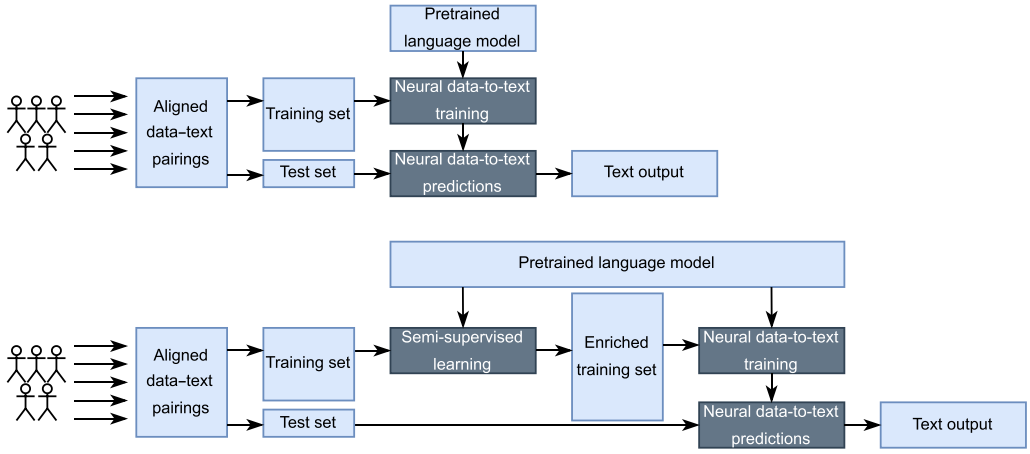

Figure 1
Schematic overview of supervised learning (top) vs. semi-supervised learning (bottom) for the current study.

that is available (Sun et al. 2020). These language models have also been found to have a beneficial impact on output quality for the data-to-text task, just like semi-supervised learning (e.g., Kale and Rastogi 2020). Both of these approaches aim to improve the data-to-text training model by providing extra information in addition to the information that is present in the training set as an enriched training set. Therefore, it is conceivable that the massive amounts of information already incorporated in a language model makes the use of semi-supervised learning redundant.

However, studies have also shown that the beneficial effect of a language model decreases when a model is overfit too much during finetuning (Greco et al. 2019). Overfitting is also more likely to occur with small datasets. The fact that semi-supervised learning increases the dataset size may therefore help against overfitting. If so, language models and semi-supervised learning would be complementary approaches that might lead to better performance when used in conjunction. However, not much is known about the effect that this combination of language-models with semi-supervised learning has in a data-to-text setting. Furthermore, provided that the addition of semi-supervised learning is beneficial, the type of semi-supervision approach that leads to the best results is currently unknown, as experimental comparisons of this are scarce as well.

This study will investigate when and how semi-supervised learning affects the output diversity and quality when used in combination with language models for data-to-text generation. Two different semi-supervised learning approaches are investigated—both utilizing pretrained Transformers models (Vaswani et al. 2017)—and the impact that these approaches may have on the output quality and diversity of a neural data-to-text system with a language model. The semi-supervised learning approaches used in this study are: (I) a **data augmentation** approach, where several variants of a training text are generated by replacing certain words with synonyms or semantically similar words, and (II) a **pseudo-label** approach, where unlabeled texts are given data labels by an information extraction (semantic parsing) model trained on the existing labeled training data. The synthetic data–text pairs obtained via these two approaches are then added to the original training set in a neural data-to-text system to generate new texts.

Even though a handful of recent papers have started exploring the benefits of semi-supervised learning for data-to-text generation, this is the first study presenting a detailed, large scale analysis (1) of different corpora, with vastly different characteristics, in two different languages (vs. previous approaches that mostly focused on only one corpus), (2) that systematically compares different methods for semi-supervised learning, also in combination with pretrained language models (vs. previous approaches that mostly did not incorporate pretrained language models in the semi-supervised learning approaches, nor the baseline), and (3) that performs an exhaustive evaluation of the different methods, by combining automatic analyses with a human evaluation, error analysis, and qualitative analysis, in line with recent best practices for evaluation (vs. previous approaches that mostly focus on automatic metrics or only conduct a limited human evaluation).

1.1 Hypotheses

Based on previous studies, we formulated several hypotheses and a more exploratory research question before conducting the study. All these hypotheses and research questions have also been preregistered before conducting the study at `https://aspredicted.org/in665.pdf`, following the advice given by van Miltenburg, van der Lee, and Krahmer (2021). In this subsection, we will state these hypotheses and research questions, followed by the rationale for the expected effects.

H1. *Extending the training set with semi-supervised learning increases the output* ***quality*** *of a neural data-to-text system with a language model (compared to a data-to-text system with a language model only trained on the base training set).*

Previous studies have consistently found that semi-supervised learning leads to improvements in output quality (e.g., Kulhánek et al. 2021), Riabi et al. 2021), Tandon et al. 2018), Alberti et al. 2019), Chang et al. 2021), Kedzie and McKeown 2019). Additionally, Sun et al. (2020) examined the dynamics between semi-supervised learning and language models in a text classification setting and found that a combination of the two approaches led to the highest classification scores.

H2. *Extending the training set using semi-supervised learning increases the output* ***diversity*** *of a neural data-to-text system with a language model (compared to a data-to-text system with a language model only trained on the base training set).*

Finetuning a text generation system with language models on small or non-diverse training data may lead to limited diversity in the output, even though language models themselves are trained on enormous amounts of text. This has to do with the known propensity for catastrophic forgetting that neural networks display (Greco et al. 2019): The model is overfit too tightly during finetuning, which leads to "forgetting" about the capabilities of language modeling. Therefore, extending the finetuning dataset with more various language using semi-supervised learning could have a positive impact on diversity, which has also been found in a study by Kulhánek et al. (2021).

RQ1. *Do the data augmentation and pseudo-label approaches differ in terms of output quality and output diversity when used as semi-supervised learning approaches in a neural data-to-text system with a language model?*

Although different semi-supervised learning approaches have shown their potential for various NLP/NLG tasks, not much is known about *which* semi-supervised learning approaches are the most effective (Sun et al. 2020).

H3. *The beneficial effect of semi-supervised learning for a neural data-to-text system with a language model on output **quality** is bigger for a small-scale dataset (CACAPO) than for a large-scale dataset (WebNLG and E2E).*

For datasets, we make a distinction between crowdsourced on the one hand (e.g., WebNLG, E2E, ToTTo; Novikova, Dušek, and Rieser 2017; Gardent et al. 2017a, 2017b; Parikh et al. 2020), and datasets that are created from "naturally occurring" human-written texts on the other (e.g., YelpNLG, RotoWire, CACAPO; Oraby et al. 2019; Wiseman, Shieber, and Rush 2017; van der Lee et al. 2020). For the construction of a crowdsourced dataset, crowdsource workers write corresponding texts for a given set of meaning representations. This approach is reasonable for the construction of a large-scale dataset, provided that time and resources are not an issue.

However, the very procedure of using crowdsource workers to verbalize a set of meaning representations without any given context ensures that texts are mostly focused on high fidelity, while sacrificing on criteria like fluency and enjoyability (van der Lee et al. 2020). This means that output from systems trained on crowdsourced datasets is less likely to contain diverse language. Alternatively, it is possible to construct a dataset with texts written in real-world scenarios, rather than lab-setting verbalizations by crowdsource workers. This can be done by collecting texts and corresponding data on a large scale, without having detail-level alignment information (as is done in Wiseman, Shieber, and Rush 2017; Wang 2019; Puduppully, Dong, and Lapata 2019), or by manually extracting aligned data from texts (as is done in Oraby et al. 2019; van der Lee et al. 2020). While such an approach is likely to facilitate more diverse and fluent language, the difficulty of this alignment task makes large-scale dataset collection a daunting endeavor.

In the current study we will compare the two most widely used data-to-text datasets: E2E (Novikova, Dušek, and Rieser 2017) and WebNLG (Gardent et al. 2017a, 2017b), which are both large-scale and crowdsource-based, and CACAPO (van der Lee et al. 2020), which is smaller-scale and based on real-world texts, in two languages (viz. English and Dutch). Previous studies have suggested that the benefits of semi-supervised learning are greater in more low-resource scenarios (Chang, Demberg, and Marin 2021).

H4. *The beneficial effect of semi-supervised learning for a neural data-to-text system with a language model on output **diversity** is bigger for a crowdsourced dataset (WebNLG and E2E) compared to a dataset based on real-word texts (CACAPO).*

Focusing on diversity, datasets based on real-world texts generally contain more diverse language than crowdsourced datasets (van der Lee et al. 2020). This implies that crowdsourced datasets have more to gain from semi-supervised learning approaches' potential to introduce more language diversity to the training set of crowdsourced datasets.

H5. *The beneficial effect of semi-supervised learning is greater when trained on a Dutch dataset (CACAPO Dutch) than an English dataset (WebNLG, E2E, and CACAPO English) for a neural data-to-text system with a pretrained language model.*

It is a well-known fact that the majority of NLP developments are focused on English, with many other languages lagging behind in terms of support (Bender et al. 2021; Riabi et al. 2021). This is also the case for language models. Although such language models exist for other languages, and multilingual variants of state-of-the-art models exist (e.g., mBert, mT5; Devlin et al. 2019; Xue et al. 2021), the size of these models is generally much smaller (Bender et al. 2021), and they are oftentimes missing

functionality. For instance, mT5 was not pretrained on downstream tasks like T5 was (Xue et al. 2021). This means that the benefits of language models in data-to-text generation is likely smaller for underrepresented languages, which in turn might mean that the beneficial effects of semi-supervised learning is greater, especially when we take into consideration that the size of non-English datasets is generally smaller as well (Riabi et al. 2021).

2. Background

2.1 Semi-Supervised Learning

The goal of semi-supervised learning is to train a model (partially) on synthetic data (as opposed to data created by humans), which may lead to a better trained machine learning model, and hence improved performance. Figure 1 gives a schematic overview of semi-supervised learning and how it differs from supervised learning, which is the standard for neural NLG. This approach has steadily grown in popularity with the rise of data-hungry neural models, and is considered especially useful when the (labeled) training set is small-scale. Within the NLG domain, we have seen applications of this approach in, for instance, question answering (e.g., Alberti et al. 2019; Riabi et al. 2021), and text simplification (e.g., Surya et al. 2019; Zhao et al. 2020).

The semi-supervised approach has also gained traction in the context of data-to-text generation. This is mostly in the form of joint learning systems, where an NLG system (that converts meaning representations into text), and a Natural Language Understanding system (that converts text into meaning representations) are feeding each other more synthetic data in a loop. Such an architecture allows for both unlabeled texts as well as meaning representations without aligned text to be included into the training data. Some studies suggest that the use of a joint learning system led to improvements on various metrics compared to a supervised NLG system (e.g., Qader, Portet, and Labbé 2019; Schmitt et al. 2020; Su, Huang, and Chen 2020; Tseng et al. 2020; Chang, Demberg, and Marin 2021).

The architecture of the current research differs from these previous studies as it utilizes unlabeled/unaligned data in a non-joint way. This approach is based on the assumption that it is easier to extract information from a text than to generate text that accurately represents information (Wiseman, Shieber, and Rush 2017). It should also be noted that previous data-to-text studies using semi-supervised approaches partition a segment of the dataset and detach the data from the texts to create unlabeled data, or they use value swapping (i.e., pairing each key with a randomly sampled value collected from the set of all data samples to obtain new combinations of key-value pairs) to create extra unlabeled data.

Our approach tries to more closely emulate the application of this task in a real-world setting by collecting and utilizing unaligned texts that are not found in the datasets. Furthermore, none of the previous studies used language models in the architecture of their data-to-text generation system and only one previous study utilized these language models for semi-supervised learning (Chang et al. 2021).

2.2 Language Models

Transformers-based language models have particularly shown their viability for generation tasks that involve meaning manipulation (e.g., summarization, text simplification,

and question answering), but studies also suggest that the inclusion of Transformers-based language models can lead to improvements in output quality for the data-to-text generation task (Chen et al. 2020; Kale and Rastogi 2020; Mager et al. 2020; Ribeiro et al. 2021). Furthermore, Transformers-based language models have been found to perform well on very small datasets, with examples existing of few-shot, one-shot, or even zero-shot learning (Brown et al. 2020). This impressive performance might suggest that including language models in the architecture of an NLG system might make semi-supervised learning approaches redundant. However, it should be noted that language models and most semi-supervised learning approaches utilize different types of data. While language models leverage data from an immense variety of different domains, semi-supervised learning approaches are generally focused on using in-domain data (Sun et al. 2020).

Therefore, a combination of language models and semi-supervised learning approaches might enhance performance, rather than canceling each other's improvements out. Sun et al. (2020) find support for this notion in a text classification context. They found that the largest performance gain was achieved when the two were combined (Sun et al. 2020). The authors also compared different semi-supervised learning approaches (in-domain pretraining and pseudo-labeling) and found performance differences between the two. Similarly, the current study also compares the performance of two semi-supervised learning approaches, but in a data-to-text generation context: **pseudo-labeling** and **data augmentation**.

2.3 Pseudo-Labeling

One of the most common semi-supervised learning approaches is the pseudo-labeling approach, where unlabeled data is assigned labels by a model, thus forming a large labeled dataset that can be used to train a model. In the context of NLP, this task is equal to information extraction (also known as semantic parsing, or natural language understanding), where a meaning representation is parsed from a text. Most of the existing semi-supervised NLG systems have utilized information extraction for the creation of synthetic training data (e.g., Qader, Portet, and Labbé 2019; Schmitt et al. 2020; Su, Huang, and Chen 2020; Tseng et al. 2020; Chang, Demberg, and Marin 2021).

However, most of these studies also apply pseudo-labeling without utilizing any language model. Besides their suitability for various generation tasks, Transformers-based language models have also shown their potential for information extraction. For example, various authors have applied BERT-based information extraction successfully on small datasets (e.g., Nguyen et al. 2019; Zhang et al. 2020a), and one of the best performers on the semantic parsing subtask of the WebNLG+ Challenge 2020 (Castro Ferreira et al. 2020) was a parser that used T5 as a pretrained model (Agarwal et al. 2020). Building on these previous findings, the current study also utilizes an approach to pseudo-labeling that includes a pretrained model.

2.4 Data Augmentation

Data augmentation generally refers to all strategies that increase training examples without explicitly collecting new data (Feng et al. 2021). This can be done for instance by adding slightly edited copies of existing texts, which has been applied in the data-to-text generation context through methods such as (back) translation (Kulhánek et al. 2021; Riabi et al. 2021).

More common, however, is data augmentation in the form of self-training, where a first version of a data-to-text model generates new texts from inputs, which are then fed to the model for further training (Heidari et al. 2021; He et al. 2020). In a data-to-text generation context, this has specifically been done through text generation from modified meaning representations (e.g., Tandon et al. 2018; Alberti et al. 2019; Chang et al. 2021), by noise injection (Kedzie and McKeown 2019), or by simply generating texts from unpaired inputs (e.g., Heidari et al. 2021; Jolly et al. 2022; Mehta et al. 2022). With few exceptions (i.e., Jolly et al. 2022), these approaches were all found to increase output quality.

A less common way of data augmentation in data-to-text generation is by using synonym replacement and text editing, which has effectively been applied for text classification (e.g., Zhang, Zhao, and LeCun 2015), and hate speech detection (Emmery et al. 2022; Rizos, Hemker, and Schuller 2019). However, recent advances in learning-based quality estimation metrics, such as RUSE (Shimanaka, Kajiwara, and Komachi 2018), BERTScore (Zhang et al. 2020b), MoverScore (Zhao et al. 2019), and BLEURT (Sellam, Das, and Parikh 2020) try to gauge the semantic similarity of generated sentences compared to a gold standard using language models. The ability of these metrics to detect synonyms and semantically similar language does illustrate the viability of synonym replacement in (data-to-text) NLG as well, as instilled knowledge of semantically similar words and phrases is the most important part of data augmentation based on synonym replacement. In the current study, we will further investigate the potential for (the synonym/semantically similar replacement approach for) data augmentation using language models as a semi-supervised learning approach in data-to-text generation.

3. Approach

3.1 Datasets

For the data augmentation approach, it is beneficial to be able to locate the exact position where data was verbalized in the text, so that augmentations in the text that have to do with the data could easily be traced back and changed in the data as well. This ensures that augmented variants of texts also align with its data counterpart. Therefore, datasets were chosen for this experiment that included such enriched information. These are: (enriched) E2E (Castro Ferreira et al. 2021), (enriched) WebNLG (Castro Ferreira et al. 2018), and CACAPO (van der Lee et al. 2020). We used the original train/development/test splits for these corpora. Specific characteristics of the datasets are discussed in more detail below.

It should also be noted that we differentiate between domains for the CACAPO and WebNLG dataset (E2E is only one domain), as we believe that treating them separately will not only result in higher performance, but also provide richer and more detailed information about the performance of the various methods. Domains in CACAPO and WebNLG are inherently different due to imbalances that exist in the data (mostly in WebNLG) and due to the very nature of the reports. More specifically: The differences in the richness of information they provide, and the complexity and diversity of the language that is used. To fully capture the effects that these domain differences have on the performance of different approaches, it is necessary to look at the sub-corpora.

Furthermore, we added synthetic data to the original dataset in various quantities. This was done to exploratively study the effects of semi-supervised learning in a more detailed fashion. It could, for instance, reveal that there is a saturation point

where adding more synthetic data stops improving performance, or that performance decreases when more synthetic data is added which indicates cascading of errors. Sizes and statistics are described in more detail below.

3.1.1 E2E (Novikova, Dušek, and Rieser 2017). E2E is focused on the restaurant domain and contains English verbalizations of data, which were collected using crowdsourcing. The data for this dataset is stored in a key-value format, similar to CACAPO. The dataset is split in a training, development, and test ratio of 76.5%/8.5%/15%, respectively. Of the three datasets in this study, E2E is the largest in terms of sheer size: Table 1 summarizes the basic statistics for E2E (and the other corpora used in this study). E2E contains 42,061 instances (i.e., aligned data–text pairs), and 840,760 tokens. Furthermore, it contains 4,862 unique meaning representations (i.e., data elements), less than the other two corpora, which suggests that this dataset contains more repetition compared with WebNLG and CACAPO.

As there is no information available on the origins of the data used for E2E, it is difficult to collect new data or find comparable restaurant descriptions online. Therefore, we used E2E+ (Roberti et al. 2020) as extra data for the pseudo-labeling approach. E2E+ is a modified version of E2E where all slot data is replaced with comparable data. For instance, *food* data is replaced using the adjectival forms of countries and nations found on Wikipedia, and *name* and *near* are replaced with New York restaurant names found

Table 1
Size-related descriptives for the standard and semi-supervised training sets.

CACAPO Dutch	**No. of instances**	**No. of unique MRs**	**No. of tokens**	**CACAPO English**	**No. of instances**	**No. of unique MRs**	**No. of tokens**
No Extension	7,367	6,590	110,391	No Extension	7,923	7,613	153,663
Dat_Aug (S)	14,719	13,258	220,753	Dat_Aug (S)	15,822	14,962	308,655
Dat_Aug (M)	22,067	19,955	331,092	Dat_Aug (M)	23,718	22,306	463,641
Dat_Aug (L)	44,067	40,072	661,943	Dat_Aug (L)	47,391	44,230	928,203
Dat_Aug (XL)	80,537	73,272	1,212,192	Dat_Aug (XL)	86,648	79,595	1,700,241
Pseu_Lab (S)	13,626	11,727	201,894	Pseu_Lab (S)	17,482	16,671	354,748
Pseu_Lab (M)	20,010	16,779	296,171	Pseu_Lab (M)	27,221	25,010	555,519
Pseu_Lab (L)	32,465	25,778	479,852	Pseu_Lab (L)	47,112	41,863	961,614
Pseu_Lab (XL)	57,251	42,286	844,537	Pseu_Lab (XL)	82,528	70,884	1,681,230
WebNLG	**No. of instances**	**No. of unique MRs**	**No. of tokens**	**E2E**	**No. of instances**	**No. of unique MRs**	**No. of tokens**
No Extension	24,404	10,672	349,712	No Extension	42,061	4,862	840,760
Dat_Aug (S)	48,732	32,086	690,545	Dat_Aug (S)	83,853	43,555	1,683,644
Dat_Aug (M)	73,059	52,225	1,031,404	Dat_Aug (M)	125,644	81,708	2,526,518
Dat_Aug (L)	146,018	112,625	2,053,905	Dat_Aug (L)	251,017	196,285	5,054,910
Dat_Aug (XL)	267,478	212,312	3,757,401	Dat_Aug (XL)	459,956	385,975	9,267,858
Pseu_Lab (S)	29,395	15,586	449,320	Pseu_Lab (S)	48,489	10,814	973,220
Pseu_Lab (M)	34,310	20,336	548,731	Pseu_Lab (M)	54,917	16,065	1,105,611
Pseu_Lab (L)	44,073	29,729	743,161	Pseu_Lab (L)	67,774	25,091	1,371,851
Pseu_Lab (XL)	62,393	47,199	1,108,450	Pseu_Lab (XL)	93,487	39,622	1,901,112

in the Entree dataset (Burke, Hammond, and Yound 1997) (e.g., *Blue Spice serves highly rated Chinese food.* becomes *El Charro serves highly rated Timorese food*). To investigate the effect of the pseudo-labeling approach, four sizes of aligned data–text information were added to the original training set: Small, medium, large, extra large. These contained 12.5%, 25%, 50%, and 100% of the E2E+ data, respectively, in the case of the pseudo-labeling approach.

3.1.2 WebNLG (Gardent et al. 2017a, 2017b). WebNLG is collected in a similar crowd-sourced manner as E2E and is derived from DBPedia properties. These properties are different from E2E and CACAPO data as they are not stored in a key-value format, but as SVO-triples (subject-verb-object). Each of these properties is related to a particular category in DBPedia. For the enriched WebNLG dataset, these domains are: `Airport`, `Astronaut`, `Building`, `City`, `ComicsCharacter`, `Food`, `Monument`, `SportsTeam`, `University`, and `WrittenWork`. While the dataset is smaller than E2E in terms of tokens and instances (24,404 instances; 349,712 tokens), it does seem more varied in its composition, as evidenced by the number of unique meaning representations (10,672) (see Table 1). Furthermore, the dataset is split by a 60%/20%/20% training, development, and test ratio.

Following Montella et al. (2020), we collected Wikipedia texts as extra data for the pseudo-labeling approach for WebNLG. These texts are similar in nature, as Wikipedia pages are generally well-connected to the DBPedia variant of the page. Furthermore, Wikipedia texts are freely available and relatively easy to collect. For each of the DBPedia categories in the WebNLG dataset, we searched for similar overview pages on Wikipedia, and then scraped all the pages in the overview or in the subcategories of the overview.[1] Then, the summary (i.e., the first paragraph of the article) was taken from each page, split on a sentence-level, labeled, and added as extra data. The training set was extended with the sentences of 125, 250, 500, and 1,000 summaries for, respectively, small, medium, large, and extra large. Because the pseudo-labeling data was split on a sentence-level, the original training set was also split on a sentence level to ensure consistency between the original training data and the input derived from the pseudo-labeling approach.

3.1.3 CACAPO (van der Lee et al. 2020). The CACAPO dataset (van der Lee et al. 2020) contains texts from the `Sports`, `Weather`, `Stocks`, and `Incidents` domain for both Dutch and English. Each domain contains information for 200 texts (1,600 texts total), paired with manually annotated data for each sentence in a key-value format. It is split up in a 76.5%/8.5%/15% training, development, and test ratio, similar to E2E.

Besides language differences, there are also topical differences between the English and Dutch part of the dataset: The `Weather` and `Stocks` report are relatively similar in their content, but the Dutch version of the `Sports` domain contains soccer reports, whereas the English version contains baseball reports (based on Puduppully, Dong, and Lapata 2019). Similarly, the Dutch `Incidents` domain contains reports about traffic incidents (from Hendriks 2019), while the English `Incidents` domain contains reports about firearm incidents (see van der Lee et al. 2020, for a detailed description). Size-wise, both the Dutch and English datasets are the smallest in this study in terms of instances (Dutch: 7,367, English: 7,923) and tokens (Dutch: 110,391, English: 153,663) (see Table 1).

1 The full list of pages that were collected can be found at `https://github.com/TallChris91/Neural-Data-to-Text-Small-Datasets`.

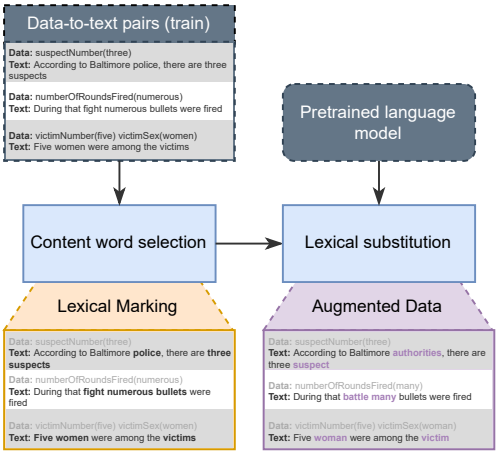

Figure 2
Pipeline for data-to-text data augmentation.

The large number of meaning representations (Dutch: 6,590, English: 7,613) indicates a relatively large variation for its size.

Unlabeled texts for the pseudo-labeling approach were scraped using the same text collection methods as were used for the CACAPO dataset. This means that human-written texts were collected from the same selection of Web sites as were used for CACAPO. Furthermore, the texts were collected using an automatic scraper, or a tool that made saving texts in a correct format as effortless as possible, as was also done in the construction of CACAPO. Similar to WebNLG, the small version of the training set was extended with the sentences of 125 articles for small, 250 articles for medium, 500 articles for large, and 1,000 articles for extra large.

3.2 Data Augmentation

For data augmentation, we use lexical substitution (McCarthy and Navigli 2007); that is, for specific words in the input (the target words) we determine multiple alternatives that are semantically similar (the substitution candidates). For this purpose, we use Emmery, Kádár, and Chrupała's (2021) implementation of Zhou et al.'s (2019) work. Under the framework of masked language modeling, to predict synonyms rather than any word fitting a particular (masked) position, Zhou et al. (2019) proposed using Dropout (Srivastava et al. 2014). The pipeline of this architecture is displayed in Figure 2.

Instead of masking a selected word in the text, Dropout is applied to the BERT-internal embedding of that word. The intuition is that, rather than BERT predicting identical words when the original word's embedding is passed, the partly-zeroed embedding results produces synonyms instead. These are then the substitution candidates, which we contextually re-rank using a similarity score.[2]

Candidates are removed if they do not match certain criteria: Their similarity scores should be > 0.9, and should not be punctuation or single characters, UNK tokens, plurals or capitalized versions of, or equal to the target word, subwords, or already exist

2 In the original work, this is a subcomponent of the ranking function. We observed little difference in ranking by adding the word probability and α weightings (which do add computational complexity).

in the sentence. BERT-large (Devlin et al. 2019) was used to generate the candidates for English, BERTje (de Vries et al. 2019) for Dutch, and Dropout was set to 0.2.

For the target words, we chose all nouns, adjectives, adverbs, and numerals—tagged using SpaCy (Honnibal and Montani 2017). Similar to Emmery et al. (2022), we fill each position with a candidate simultaneously (i.e., using the highest ranked candidates for each target word to produce the first augmented instance, and so on; e.g., *"What will the **weather** be like this **afternoon** in **Preston**?"* → *"What will the **air** be like this **evening** in **Manchester**?"*). We repeat this step for a maximum of twenty instances. If target words do not have up to the maximum amount of substitution candidates, they are left as the original words instead. The top 1 (small), 2 (medium), 5 (large), and 10 (extra large) instances of each text, based on their BERT similarity score with respect to the original sentence, were then added to the training datasets. As previously noted, enriched versions of corpora were used for data augmentation to ensure that augmentations were also applied to the aligned data.

We acknowledge the difficulty of measuring the performance of data augmentation using automatic metrics, since most metrics are based on a comparison to a gold standard. Furthermore, language model-based semantic distance metrics (such as BERTScore and BLEURT) are very similar in nature compared to the data augmentation approach used in this study, which might make their scores more akin to a manipulation check rather than an accurate reflection of semantic similarity between an augmented sentence and its original. Still, performing an evaluation using these metrics offers novel information as it measures sentence-level semantic consistency, which has not been measured in full during the data augmentation process.

Therefore, we calculated the average BLEU (Papineni et al. 2002), BLEURT (Sellam, Das, and Parikh 2020), and BERTScore (Zhang et al. 2020b) scores of the augmented texts compared to their original. A lower BLEU score (a straightforward metric that measures text-overlap between a candidate and reference) and higher performance on BLEURT and BERTScore (metrics that aim to measure semantic similarity between a candidate and reference) might suggest that texts have been augmented fundamentally, while still conveying a semantically similar message. Furthermore, we used LanguageTool[3] to calculate the difference in grammatical errors compared to the original sentences, as an indicator for (relative) grammatical correctness.

The results indicate that the data augmentation was generally effective. The BLEU scores are mostly around 10–20 for all domains, although a few exceptions exist. These low BLEU scores suggest that a large chunk of the original sentences were modified, meaning that the training data became more varied. We also see that the BLEURT and BERTScore numbers are higher than BLEU for almost every domain, with scores in the 40–50 range for all domains. The BERTScores for the Dutch domains are an exception, which rise above 80 due to the fact that rescaling with a baseline is not possible for this language.

Nevertheless, these scores seem to indicate that the semantic similarity to the original text is kept relatively intact. Finally, the difference in grammatical errors is small, with generally only 0.1 to 0.2 more errors being found in the augmented texts compared to the original texts. This suggests that data augmentation adds few new grammatical errors to the texts. Thus, overall, we can see these scores as an indicator that the data augmentation approach indeed manages to modify a sentence fundamentally, while still

3 https://languagetool.org.

Table 2
BLEU, BLEURT, BERTScore, and mean difference in grammatical errors; comparing the original texts to the top 10 (XL-size) augmented sentences.

Dataset	Domain	BLEU	BLEURT	BERTScore	Δ Grammar
CACAPO	Incidents (EN)	17.32	45.97	51.49	+0.14
	Sports (EN)	40.30	43.05	47.09	+0.15
	Stocks (EN)	22.83	48.41	46.92	+0.14
	Weather (EN)	30.21	47.55	45.67	+0.10
	Incidents (NL)	17.44	38.04	83.58	+0.03
	Sports (NL)	22.69	34.67	83.77	+0.11
	Stocks (NL)	20.11	35.75	84.18	-0.06
	Weather (NL)	26.78	43.63	83.11	+0.19
WebNLG	Airport	13.65	40.50	36.28	+0.12
	Astronaut	8.91	45.19	55.13	+0.07
	Building	11.39	43.63	42.15	+0.12
	City	9.26	44.23	40.69	+0.04
	ComicsCharacter	64.84	40.00	44.16	+0.06
	Food	41.11	44.06	43.90	+0.13
	Monument	16.26	42.86	47.63	+0.07
	SportsTeam	14.54	40.64	41.06	+0.10
	University	32.47	46.95	44.94	+0.08
	WrittenWork	20.86	42.84	40.39	+0.10
E2E		24.27	46.78	57.99	+0.09

keeping the text relatively semantically similar to the original and relatively error-free (see Table 2).

3.3 Pseudo-Labeling

Similar to Schmitt et al. (2020), we framed the pseudo-labeling task as a text-to-text translation task, as this approach could handle the differences in data formats between all three datasets most effortlessly and effectively (compared with, for instance, span labeling, or extractive question answering). While the most straightforward text-to-text translation purpose is to translate a text from, for instance, English to German, text-to-text translation can actually be used effectively for a multitude of natural language processing tasks, as most of these involve conversion of one text format into another. T5 (Raffel et al. 2020) was developed with this purpose in mind.

This architecture, also known as Text-to-Text Transfer Transformer (T5), is a large pretrained language model resulting from an empirical survey to determine which transfer learning techniques work best. Different from classification language models such as BERT, T5 works as a unified text-to-text-approach where all the NLP tasks are reframed such that its inputs and outputs are strings. This text-to-text framework is a multi-task one, sharing the parameters, loss function, and hyperparameters on any NLP task, including machine translation, document summarization, question answering, and classification tasks (e.g., sentiment analysis). To set the desired task for the model, a prefix needs to be inserted in the input such as "translate English to German" for the machine translation task or "summarize" for the summarization one.

Attribute	Value
Name	*Wildwood*
eatType	*pub*
food	*Indian*
area	*city centre*
familyFriendly	*no*
near	*Raja Indian Cuisine*

↓

name @SEP@ Wildwood @EOF@ eatType @SEP@ pub @EOF@ food @SEP@ Indian @EOF@ area @SEP@ city centre @EOF@ familyFriendly @SEP@ no @EOF@ near @SEP@ Raja Indian Cuisine

Figure 3
Example of attribute-value pairs and the corresponding data string. @SEP@ = seperator, @EOF@ = end of field.

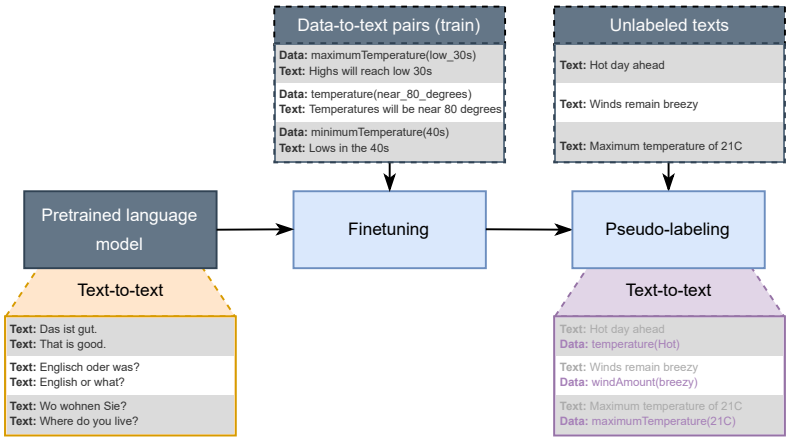

Figure 4
Pipeline for data-to-text pseudo-labeling.

In this case, we "translated" a "data language" to Dutch or English using T5-large (Raffel et al. 2020) for pretraining of the English pseudo-label model, and mT5-large (Xue et al. 2021) for pretraining of the Dutch pseudo-labeling model (following Agarwal et al. 2020). This was done using run_translation.py from `https://github.com/huggingface/transformers` (Wolf et al. 2019) with 30 epochs and a batch size of 8. T5 and mT5 were further finetuned on the original training and development set of the CACAPO, E2E, and WebNLG datasets and applied to the test set to calculate performance. Furthermore, the trained models were applied to the unlabeled new texts that were collected to extend the training set.[4]

More specifically, we converted the data into a structured string format that follows the data structure of the dataset (see Figure 3) using "translate Dutch [resp., English] to Data: [...]" as the prefix command. For the output, this string format was then converted back into structured data using a simple rule-based script. The pipeline of this architecture is displayed in Figure 4.

4 Other pretrained language models, and other approaches, have also been investigated in this experiment for pseudo-labeling as well as data augmentation and data-to-text generation. The results of these approaches are reported on the GitHub page, which will be continuously updated with the results of new methods: `https://github.com/TallChris91/Neural-Data-to-Text-Small-Datasets`.

Table 3
Precision, Recall, and F1 scores of information extracted by our pseudo-labeling system.

		Dev			Test		
		P	**R**	**F1**	**P**	**R**	**F1**
CACAPO	Incidents (NL)	83.00	85.17	84.07	74.27	78.76	76.45
	Sports (NL)	73.65	77.97	75.75	74.33	76.03	75.17
	Stocks (NL)	85.96	89.17	87.54	90.37	89.72	90.04
	Weather (NL)	81.16	87.63	84.27	85.57	89.92	87.69
	Incidents (EN)	79.63	82.43	81.01	77.85	79.84	78.83
	Sports (EN)	79.02	80.07	79.54	79.95	79.95	79.95
	Stocks (EN)	80.60	84.84	82.66	83.08	79.27	81.13
	Weather (EN)	83.61	80.94	82.25	79.83	82.77	81.27
WebNLG	Airport	89.70	88.76	89.23	91.03	89.85	90.43
	Astronaut	96.13	95.41	95.77	97.00	94.95	95.97
	Building	89.81	91.04	90.42	89.86	89.31	89.59
	City	73.61	73.61	73.61	63.02	28.63	39.37
	ComicsCharacter	95.96	98.17	97.05	96.53	95.59	96.06
	Food	87.67	88.56	88.11	89.02	88.39	88.70
	Monument	72.38	68.88	70.59	52.83	50.76	51.77
	SportsTeam	81.52	81.79	81.65	88.07	88.22	88.15
	University	95.52	93.43	94.46	93.20	91.13	92.15
	WrittenWork	93.35	93.23	93.29	92.89	91.48	92.18
E2E		85.65	91.74	88.59	88.17	82.72	85.36

We evaluated the performance of the pseudo-labeling approach by calculating the precision, recall, and micro-averaged F1 score on the development and test sets of all datasets. While we believe that these measures give a robust indication of the labeling quality, it should be noted that the model might not generalize well to the unlabeled texts, especially when the unlabeled texts are highly dissimilar from the texts seen in training (for instance, the pseudo-labeling model trained on E2E showed a considerable drop-off on the E2E+ data.[5])

Overall, the scores indicate that this pseudo-labeling approach performs well, with F1 scores well above the 70s for CACAPO and even in the high 80s and 90s for WebNLG and E2E (see Table 3). Two notable exceptions are the `City` and `Monument` domains for WebNLG, which achieve much lower scores than other domains. This is likely caused by imbalanced data in the WebNLG dataset, which is especially prevalent in the `City` and `Monument` domain.

3.4 Data-to-Text Generation

The data-to-text approach utilized in this study was a neural end-to-end architecture where a set of input data is directly converted into English or Dutch text. This was done using Any2Some,[6] which uses language models from the HuggingFace API (Wolf et al.

5 On the synthetic data, P 54.07, R 60.93, and F1 57.30 was achieved.
6 https://github.com/ThiagoCF05/Any2Some.

2019) to perform data-to-text generation, while offering advantages such as automatically clustering verbalizations based on the same data. Similar to the pseudo-labeling step, we used T5-large (Raffel et al. 2020) for the data-to-text conversion step as well. This time, using the text-to-text nature of the language model to perform data-to-text generation using "Verbalize: [...]" as the prefix command.

As mentioned previously, T5 has been developed as an approach capable of handling a multitude of Natural Language Processing tasks where the inputs and outputs are reframed as strings. For this system, the input and output used in the pseudo-labeling step was essentially reversed: The data was again converted to a structured string format but this time used as input, with English or Dutch text serving as output. Previous research has suggested that T5 is a capable language model for the data-to-text generation task (Kale and Rastogi 2020).

The model was finetuned on all individual domains for 16 epochs with a learning rate of 1e-5, early stopping of patience 5, and batch size of 2. Input and output strings were trimmed to a maximum size of 180 sub-tokens. For the Dutch CACAPO domains we used mT5-large, with the same hyperparameters except for 50 epochs. More epochs were necessary for this model to be properly trained, as mT5 was not trained on downstream tasks. Some examples of the input and output of the data-to-text generation system for each dataset and semi-supervised learning method can be found in Table A.1.

4. Evaluation

The goal of this evaluation study was to investigate the contribution of the semi-supervised learning methods in data-to-text NLG. To investigate this, the evaluation study consisted of three parts: an automatic evaluation, a quantitative human evaluation, and an error analysis.[7] We aimed to follow the best practice guidelines as described in van der Lee et al. (2021) as much as possible in the setup and reporting of the evaluation study.

For the automatic evaluation, multiple metrics were used to estimate output quality and output diversity. The quantitative human evaluation experiment measured aspects of text quality to further determine the performance of the different semi-supervised approaches relative to each other, and finally an error analysis was performed on the 15 worst scoring sentences per *dataset* × *semi-supervised learning approach* combination to investigate the shortcomings and challenges for each semi-supervised learning approach.

4.1 Automatic Evaluation

The performance of the three types of semi-supervised learning that were investigated in this research (no extension, data augmentation, and pseudo-labeling) was first tested for all domains in the CACAPO, E2E, and WebNLG datasets using automatic metrics that measure text quality and diversity. The text quality metrics served as a first test for H1, H3, H5, and RQ1. The text quality metrics employed in this study are displayed in Table 4.

7 Ethics clearance was obtained from the `Tilburg University School of Humanities and Digital Sciences Research Ethics and Data Management Committee` for this experiment (code: 2019.40). Furthermore, the study was pregistered at `https://aspredicted.org/in665.pdf` and the results of this study are available via `https://figshare.com/s/3959076f2d69d1381ccc`.

Table 4
Definitions of the automatic metrics for text quality used in this study.

BLEU	Measures exact word match precision between model output and one or more references.
NIST (Doddington 2002)	Similar to (corpus) BLEU, but adds more weight to more rare words.
METEOR (Banerjee and Lavie 2005)	Measures precision and recall of exact word matches between a reference and a candidate, also adds stemming and synonym matching.
ROUGE-L (Lin 2004)	Looks at the Longest Common Subsequence between model output and one or more references and calculates the F1 score.
BERTScore	Measures the F1 score or the similarity between model output and one or more references, instead of exact matches, it computes similarity using contextual embeddings.

Table 5
Definitions of the automatic metrics for text diversity used in this study.

Average sentence length (ASL)	Average number of tokens per sentence.
Standard deviation of the sentence length (SDSL)	How much variation there is in the number of tokens per sentence.
Number of types (Types)	Number of unique word types in the output.
Mean segmented type-token ratio (TTR 1)	Divides the generated texts into equal segments of a given token length (here: 100 tokens) and calculates the average type-token ratio of all these segments.
Bigram TTR (TTR2)	Average type-token ratio of bigram types per 100 bigram tokens.
Percentage of novel texts (%Novel)	Texts generated by the system that do not occur in the training and development data.
Coverage (Cov)	The percentage of learnable words (i.e., words in the original training or development set) that are recalled in the generated output.
Novelty (Nov)	The percentage of novel words (i.e., words that do not appear in the original training or development set) that are in the generated output.
Local Recall (Loc1)	The percentage of important words (i.e., adjectives, verbs, nouns, and adverbs) in a given test set text that are recalled by the system's generated text.

Furthermore, we use the diversity metrics based on van Miltenburg, Elliott, and Vossen (2018). These metrics are used to test H2 and H4, as they provide an objective and complete image of the diversity in the output of the systems, which cannot be measured as accurately with sentence/phrase-level human evaluation. The diversity metrics used in this study are described in Table 5.

4.2 Quantitative Human Evaluation

4.2.1 Participants. Participants of this study were recruited via Prolific, a crowdsourcing platform. For participation, participants received $4.80 (the recommended amount according to the platform). In total, 193 people participated in the study, which was divided up in a Dutch version and an English version. In the recruitment phase, only participants who were native Dutch located in the Netherlands were recruited for the Dutch version, and native English speakers located in the United States for the English version. This resulted in 41 participants in the Dutch version, of which the majority

were men (56%) between the ages of 18 and 34 years (90%). Furthermore, the majority of the Dutch sample had a university (of applied sciences) degree (79%). For the English version, 152 people participated in the study. The majority were women (64%), roughly half of the participants were between the ages of 18 and 34 years (48%), and the majority had attended or completed college (87%).

4.2.2 Design. To ensure that we captured the variety found in all datasets and among all semi-supervised learning approaches, we measured the text quality of outputs from the 3 investigated semi-supervised learning approaches on all 19 domains in the datasets we used (CACAPO: 8, E2E: 1, WebNLG: 10). We randomly sampled a total of 40 items per semi-supervised learning approach–domain combination, leading to a total of $19 \times 3 \times 40 = 2,280$ trials. Each trial was judged a total of 5 times to obtain a stable judgment of the trial.

Each participant was randomly assigned to a dataset domain. Dutch-speaking participants were randomly assigned to 1 of 4 domains (the four Dutch CACAPO domains), while English-speaking participants were randomly assigned to 1 of the other 15 (English) domains. Furthermore, each dataset domain had 2 versions, with each version containing 60 outputs total from the systems trained on the XL data (20 per semi-supervised learning approach) that were all not present in the other version. This number of outputs was chosen to ensure that the sample contained enough variety to be representative of the variation in the full dataset, while the number of stimuli presented to individual participants was still manageable for them. Participants were randomly assigned to 1 of the 2 versions. The 60 outputs were presented in random order to compensate for potential order- or fatigue effects.

4.2.3 Procedure. A survey was created using the Qualtrics platform. First, a general introduction of the experiment and a consent form was given to the participants. After consenting to participate in the research, the participants were given detailed instructions about the experiment they were about to participate in. These instructions included guidelines on how to read the data input and the output texts, and how to rate said output texts. Furthermore, definitions were given for the scales they had to rate, and examples were given about good output texts and bad output texts. Instructions, guidelines, definitions, examples, as well as the questions themselves (as shown below), were translated to Dutch for the Dutch version of the evaluation to ensure that monolingual Dutch participants were able to comprehend the contents.

After these instructions, participants were asked to provide some demographic information and then the experiment started. Participants were shown a table with the original input data accompanied by a generated text from the NLG system trained on data–text pairings from one of the three semi-supervised learning approaches (no extension, data augmentation, pseudo-labeling). A selection was made of inputs that contained between 2 and 6 data elements, to keep the input data relatively understandable for participants. The generated texts were one sentence long (for CACAPO and WebNLG), or a few sentences long (between 1 and 6; for E2E). After viewing the input data and output texts, participants were asked to rate the texts on multiple items. Definitions of each item could also be found by hovering over the item.

We measured *fluency* using four seven-point Likert scale items based on Sundar (1999) and Clerwall (2014) (consistency was high, with $\alpha = .97$). The items were introduced by "This sentence/short text is...", followed by "Clear" (The overall message of the sentence/short text is clear.), "Coherent" (It is easy to follow the connections in

the sentence/short text. The different pieces of information are connected in a correct way.), "Understandable" (The sentence/short text is written in a way that is easy to understand. There are no strange word choices or phrases that make the sentence/short text confusing.), and "Well-written" (The sentence/short text is fluent and easy to read.).

Correctness was measured using three seven-point Likert scale items based on Hoorn and van Wijngaarden (2010) (consistency was high, with $\alpha = .92$). The items were introduced by "Based on the data table, the information in this sentence/short text is...", followed by "Factual" (The sentence/short text only describes the data in the data table. There is no extra information being described in the sentence/short text that is not represented in the data table.), "Accurate" (The information in the data table is represented correctly in the sentence/short text. There are no mistakes in the names and numbers, for instance.), "Complete (All the (important) information from the data table is represented in the sentence/short text. There is no information missing in the sentence/short text that is represented in the data table.)".

Grammaticality was measured using one multiple-choice question containing 4 options, based on Ross (1979). The question was introduced by: "How grammatically correct is this sentence/short text?" followed by (1) "The sentence/short text sounds perfect. I would use it without hesitation.", (2) "The sentence/short text is less than perfect – something in it just doesn't feel comfortable. Maybe lots of people could say it, but I never feel quite comfortable with it.", (3) "Worse than 2, but not completely impossible. Maybe somebody might use the sentence/short text, but certainly not me. The sentence/short text is almost beyond hope.", (4) "The sentence/short text is absolutely out. Impossible to understand, nobody would say it. Un-English." For the results section, this domain was reverse-coded to make the scores better interpretable.

Finally, after rating all sentences/short texts, participants were fully briefed on the goal of the study, reminded of the contact addresses if they had more questions about the research, and thanked again for participation.

4.3 Error Analysis

An error analysis was performed to get a better understanding of the exact errors that can be found in the NLG output, which in turn may help to improve the various systems. The 15 worst-scoring texts (on the average of all three measured constructs) in the quantitative human evaluation experiment for each system-dataset combination (a total of 180 texts) were analyzed by 7 human annotators, all experts in language and communication and proficient in both Dutch and English; none having previously seen the output of the various systems.

All annotators coded 19 sentences jointly and 23 sentences individually. Cohen's Kappa for multiple raters (Davies and Fleiss 1982) was calculated for the jointly annotated part,[8] resulting in $\kappa = .45$. This indicates moderate agreement (Landis and Koch 1977). Eleven error categories for these 180 texts were developed based on Castro Ferreira et al. (2019), and less straightforward error categories were accompanied by a short description with examples. See Table 6 for an overview of the items. If annotators selected "Other", they were able to input text to describe the error category they felt they encountered.

8 Using NLTK's `multi_kappa` function.

Table 6
Questions asked in the error analysis study.

• Does the text contain information that is not reported in the data table?
• Is the text missing information that is in the data table?
• Did you find any mistake involving the references? (e.g., "Indian cuisine" in the data table becomes "German cuisine" in the text, or the reference is not explicitly mentioned: "He is the leader of the country." instead of "Joe Biden is the leader of the United States of America.")
• Did you find any mistake involving the verb form? (e.g., The boy "play" soccer instead of "plays", or plainly missing a verb: "The boy soccer")
• Did you find any mistake involving the determiners? (e.g., "An" boy.)
• Did you find any mistake involving the punctuation or capitalization in the sentence? (e.g., "The,boy is here;" or "the Boy Is here")
• Did you find any mistake involving strange lexical choices? (e.g., The player "shot" the goal.)
• Did you find any mistake involving illogical/unnecessary repetition of words or phrases? (e.g., "The the the" boy)
• Did you find any mistake involving connections between data points? (e.g., "The leader of the cheeseburger is Barack Obama.")
• Is the sentence (or: one of the sentences) missing important parts to make it a full sentence? (e.g., "The maximum temperature is." instead of "The maximum temperature is 12 degrees Celsius.")
• Other (specify below)

5. Results

5.1 Automatic Evaluation

5.1.1 Automated Metrics for Text Quality. The automatic analysis results are summarized in Table 7 for the overall datasets, and presented per domain in Table B.1. Additionally, Figure 5 shows the effects of increasing the synthetic data from the semi-supervised learning methods on BLEU scores. Inspection of Table 7 reveals a clear pattern: For CACAPO the pseudo-labeling approach consistently leads to the highest automatic metric scores, while for E2E and WebNLG data extension does not lead to better automatic scores overall. This pattern is consistent among all metrics, but we will zoom in on the BLEU differences in this section.

Dataset Differences. The results of the automatic metrics suggest that clear differences between datasets exist in the output quality achieved with the different semi-supervised learning approaches. For the English CACAPO dataset, BLEU scores, for instance, improved by 5.71 on average with the pseudo-labeling approach (30.50 to 36.21), but the data augmentation approach led to an average BLEU decrease of 6.13 compared with no extension (30.50 to 24.37). The positive effect on automatic metric scores was more noticeable for the Dutch part of the CACAPO dataset, where BLEU scores increased by 20.31 on average for the pseudo-labeling approach (33.94 to 54.25) and a 4.36 BLEU improvement for the data augmentation approach compared with the no extension approach (33.94 to 38.30).

The pseudo-labeling approach for WebNLG generally led to a decrease in automatic metric scores (compared with no extension) albeit relatively small, with an average BLEU decrease of 3.36. However, the BLEU decrease for the data augmentation approach (compared with no extension) was more noticeable, with a 20.20 BLEU decrease. Decreases in automatic metric scores for the semi-supervised learning approaches compared with no extension were also observed for E2E: A 15.54 BLEU score difference for pseudo-labeling and a 37.64 decrease for data augmentation (see Table 7).

Table 7
Automatic metric results of the different (XL-format) semi-supervised learning approaches (No_Wei = no model weights used, No_Ext = no training set extension, Dat_Aug = data augmentation, Pseu_Lab = pseudo-labeling, Dat + Pseu = combination of data augmentation and pseudo-labeling) for each dataset (bold = highest).

Dataset	Train type	BLEU	NIST	BERTScore	METEOR	ROUGE-L
CACAPO (en)	No_Wei	3.89	1.63	13.54	20.55	18.57
	No_Ext	30.50	6.77	59.51	56.05	51.34
	Dat_Aug	24.37	6.30	52.15	48.80	45.89
	Pseu_Lab	**36.21**	7.55	63.83	59.93	56.55
	Dat + Pseu	36.18	**7.63**	**63.97**	**60.12**	**56.41**
CACAPO (nl)	No_Wei	23.21	4.59	79.93	40.57	39.48
	No_Ext	33.94	6.77	84.60	52.91	51.97
	Dat_Aug	38.30	7.56	86.86	59.01	58.31
	Pseu_Lab	**54.25**	**9.30**	**89.84**	**68.74**	**68.05**
	Dat + Pseu	50.84	8.91	89.35	66.66	66.24
E2E	No_Wei	35.12	4.39	55.83	61.30	34.58
	No_Ext	**66.05**	**7.08**	**79.40**	**80.21**	**44.97**
	Dat_Aug	28.41	4.15	56.41	62.49	33.40
	Pseu_Lab	50.51	4.65	63.12	60.39	38.48
	Dat + Pseu	51.49	5.06	66.79	64.25	38.92
WebNLG	No_Wei	27.69	5.70	46.28	50.79	40.52
	No_Ext	**47.91**	**8.74**	**71.3**	**71.57**	**59.88**
	Dat_Aug	27.71	5.95	52.75	53.23	45.33
	Pseu_Lab	44.55	8.32	67.82	68.74	56.82
	Dat + Pseu	39.70	7.63	63.35	64.00	52.74

Moreover, Figure 5 shows that an increase in synthetic training data leads to a decrease in BLEU scores for both E2E and WebNLG. This decrease could suggest a cascading of errors where issues in the synthetic data negatively impact the quality of the generated output. Alternatively, it could be that the texts in E2E and most WebNLG domains are relatively homogeneous. Introducing more deviations from these texts increases the diversity of the output, which results in lower scores on automatic metrics that use a golden standard. This could also explain the relatively low scores of the data augmentation approach, where the quantity of the deviations may have the biggest impact on the heterogeneity of the output. This possibility is further examined with the diversity metric scores, the quantitative and qualitative human evaluation, and the error analysis.

Dataset Domain Differences. For Dutch CACAPO, the most extreme improvement was reached for the `Weather` dataset (a 52.38 improvement for pseudo-labeling compared to no extension; see Table B.1). It is possible that the small size and limited vocabulary of the original `Weather` training set was insufficient for a neural NLG system to be properly trained on, whereas it was when the extended training set from the semi-supervised learning approaches were applied. Figure 5 also shows that increasing the amount of data leads to higher BLEU scores for both Dutch and English CACAPO—that is, for the

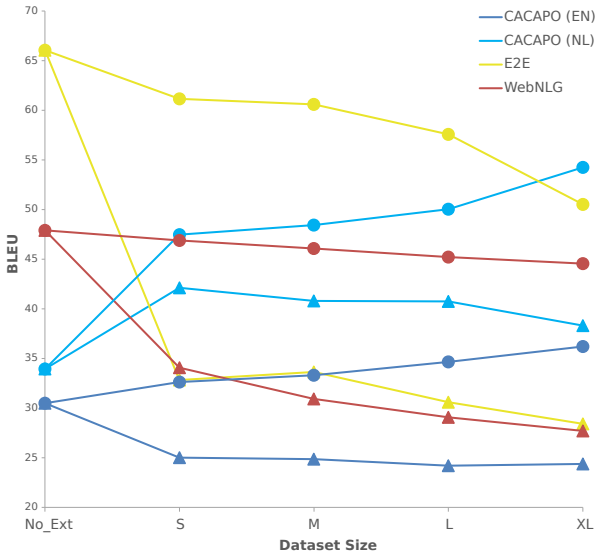

Figure 5
BLEU scores of the datasets per dataset extension. Round markers = pseudo-labeling; triangle markers = data augmentation.

pseudo-labeling method. This provides further support for the notion that extending the training set has a positive impact on output quality.

Interestingly, for WebNLG, the pseudo-labeling approach performed well for the `City` and `Monument` domains where F1 scores in the pseudo-labeling step were noticeably worse (see Table B.1). This supports the notion that semi-supervised learning is mostly effective in situations where the original training set is small (as stated in H3), filling in the training data deficit by providing more examples.

Ablation Study. An extension to the original study's design was implemented where the configuration of T5 was initialized without the weights associated with the model for the training phase. The inclusion of this model provides a better overview of the baseline's strength (i.e., the no extension model). Furthermore, a model was trained that leveraged training data enriched by both the pseudo-labeling as well as the data augmentation approach, to investigate whether a combination of the two impacts output quality.

Table 7 shows that for all datasets, a model initialized without weights performs markedly worse on all metrics compared with the models that do include the pretrained model weights. This is especially noticable for the English CACAPO dataset, where the model without weights clearly struggles. This corroborates findings by van der Lee et al. (2020), who also found that a deep learning model without pretrained data struggled with the CACAPO dataset. The characteristics of this dataset with its lack of repetition and variability make it difficult for such approaches to succeed.

Figure E.1 and Table 7 illustrate the effects of combining both data augmentation and pseudo-labeling. Generally, it performs similar to the best-performing semi-supervised learning approach, albeit a bit worse. Only for the English CACAPO dataset does the combination of the two approaches achieve higher scores, although the scores

are almost similar to those achieved with the pseudo-labeling method. Combining the two approaches therefore does not seem to benefit output quality.

Summary. In any case, while these automatic metrics do not provide all-around support for the notion that semi-supervised learning combined with a language model increases output quality compared to an NLG system with a language model that is only finetuned on the original dataset (H1), the results are in line with H3: The beneficial effect of semi-supervised learning in the text quality metrics is only noticeable for the dataset (categories) that are small-scale or unbalanced. They are also in line with H5: The highest increase of the semi-supervised learning methods was observed for the Dutch CACAPO dataset.

5.1.2 Automated Metrics for Text Diversity. A summary of the diversity metrics can be found in Table 8 on a dataset level, and on domain level in Table B.2. Furthermore, Figure 6 shows the impact of an increase in the synthetic data from the semi-supervised learning methods on coverage scores. Overall, data extensions generally lead to higher diversity scores. The semi-supervised learning approaches seem to generate more diverse output compared with no training set extension, but the semi-supervised learning approach that gives the most diverse output seems to differ per dataset (domain) and per diversity metric. Therefore, we will discuss the outcomes grouped by different diversity metrics.

Table 8
Average sentence length, standard deviation of sentence length, mean-segmented type-token ratio (TTR), bigram TTR, percentage novel descriptions, coverage, novelty and local recall with importance class 1 (bold = highest) per dataset and (XL-format) semi-supervised learning approach.

Dataset	Train type	ASL	SDSL	Types	TTR1	TTR2	%Novel	Cov	Nov	Loc1
CACAPO (en)	No_Wei	**24.51**	**25.43**	838	0.38	0.57	92.04	0.15	0.03	0.14
	No_Ext	17.26	8.29	3,502	0.66	0.93	98.24	0.58	0.20	0.53
	Dat_Aug	17.39	8.54	**3,797**	**0.68**	**0.95**	**99.80**	0.61	**0.24**	0.51
	Pseu_Lab	17.56	9.21	3,709	0.67	0.93	98.50	0.60	0.22	**0.57**
	Dat + Pseu	17.81	9.53	3,770	0.67	0.93	98.37	**0.62**	0.22	**0.57**
CACAPO (nl)	No_Wei	**16.38**	**10.41**	1,720	0.54	0.81	83.88	0.37	0.10	0.42
	No_Ext	14.65	7.70	2,748	0.58	0.85	98.86	0.56	0.20	0.52
	Dat_Aug	14.31	6.09	2,828	0.63	**0.91**	**98.93**	0.56	0.22	0.57
	Pseu_Lab	14.99	6.30	**3,176**	**0.65**	**0.91**	93.54	**0.64**	**0.24**	**0.66**
	Dat + Pseu	14.59	6.10	2,953	0.64	**0.91**	94.60	0.61	0.20	0.64
E2E	No_Wei	33.81	**30.54**	176	0.32	0.48	99.84	0.15	0.01	0.10
	No_Ext	28.58	7.66	120	0.34	0.50	**100**	0.11	0.00	**0.11**
	Dat_Aug	**34.42**	7.73	223	**0.38**	**0.55**	**100**	**0.16**	0.03	0.10
	Pseu_Lab	23.22	5.26	115	0.26	0.38	**100**	0.07	0.03	0.08
	Dat + Pseu	23.83	5.58	**226**	0.32	0.49	**100**	0.14	**0.06**	0.08
WebNLG	No_Wei	15.91	**8.51**	1,748	0.40	0.61	72.47	0.55	0.04	0.48
	No_Ext	16.01	6.71	2,136	0.43	0.68	79.78	0.69	0.02	**0.69**
	Dat_Aug	15.87	6.91	2,311	0.43	0.71	**97.71**	0.62	**0.15**	0.49
	Pseu_Lab	**16.42**	6.70	2,404	0.45	0.72	81.13	**0.73**	0.08	0.66
	Dat + Pseu	16.22	6.80	**2,527**	**0.46**	**0.74**	80.10	**0.73**	0.12	0.62

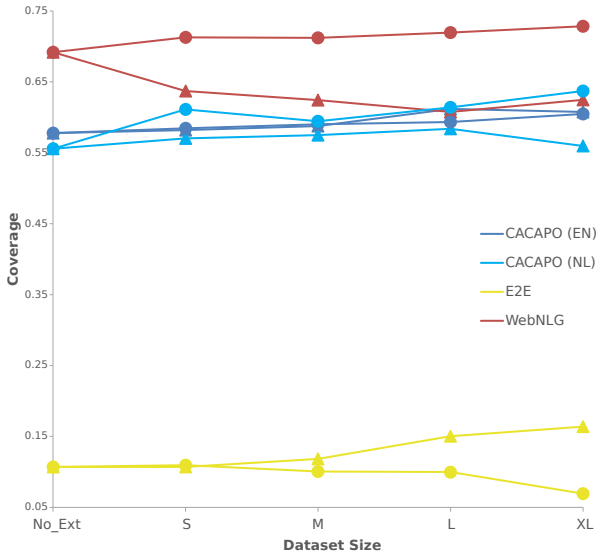

Figure 6
Coverage scores of the datasets per dataset extension. Round markers = pseudo-labeling; triangle markers = data augmentation.

Average Sentence Length, Standard Deviation of the Sentence. Average sentence length and standard deviation of the sentence can be an indicator of perceived diversity in a text: Longer sentences tend to contain more variation, and bigger differences between sentence length make the output more heterogeneous. It could be expected that the pseudo-labeling approach in particular affects sentence length standard deviation if the sentences that this approach introduces are also of a varied sentence length, while the average sentence length should not change too much for CACAPO and E2E, as the newly introduced sentences come from similar sources as the sentences in the original training set. For WebNLG, changes in average sentence length can be expected for the pseudo-labeling approach as the sentences are from a different source (Wikipedia vs. Crowdsourced). The data augmentation approach is expected to keep the standard deviation of sentence length, and average sentence length similar, as this approach perturbs words but generally keeps sentence structure and length the same.

The effect of pseudo-labeling, however, is only partially according to expectations; pseudo-labeling obtained the highest standard deviation score for 9 of the 19 domains, but only for 1 dataset overall (see Table 8; Table B.2). In terms of sentence length, it only shows a clear difference (decrease) compared with the no extension approach for E2E, and only marginal differences for the other datasets and dataset domains. Data augmentation indeed shows similar scores for average sentence length and standard deviation of sentence length consistently among datasets and dataset domains.

Number of Types, Type–Token Ratios, Percentage of Novel Texts, Novelty Score. The number of types, type-token ratios, percentage of novel texts, and the novelty score are all direct indicators of lexical diversity. For these metrics, the improvements of the semi-supervised learning approaches are also the most pronounced. The data augmentation and pseudo-labeling approach each seemed to perform best in terms of increasing lexical diversity on roughly half of the datasets and dataset domains.

For the English CACAPO, and E2E, data augmentation seemed to result in the highest lexical diversity scores, while these scores were highest for the pseudo-labeling approach in the case of Dutch CACAPO and WebNLG. For the Dutch CACAPO dataset, this may have to do with the nature of the language model used for data augmentation: BERT-large for English (Devlin et al. 2019) is likely better able to inject diverse perturbations compared with BERTje (de Vries et al. 2019), which is a Dutch translation of BERT-base.

For WebNLG it might have to do with the nature of the texts that were used for pseudo-labeling: The Wikipedia texts are probably more dissimilar (thus injecting more diversity) to the texts in the training set, compared with the pseudo-labeled texts used for the other datasets. It is worth noting that the novelty scores of the no extension approach for the WebNLG and E2E datasets are consistently close to 0, meaning that (almost) no new words have been introduced for these datasets that were not found in the training or development data. This absence of new words possibly showcases catastrophic forgetting (Greco et al. 2019) (which may occur when a training set is decently sized for finetuning), which leads to a neural NLG system forgetting about the language model's capabilities.

Also worth noting is that the percentage of novel texts shows that (almost) all generated texts in CACAPO and E2E are different from the texts in the training data, thus creating a ceiling effect for these datasets. This has to do with the setup of the test set, where input data rarely, if ever, overlaps with the input data in the training and development sets. This metric is more interesting for WebNLG, where we see that non-novel sentences are only rarely generated with the data augmentation approach, while this does occur more frequently with the pseudo-labeling approach and no extension approach.

Coverage, Local Recall. Coverage and local recall measure text retention in the output compared to the training set. Coverage is a global recall metric, measuring how many of the word types from the training and development sets were retained in the output. A higher coverage score suggests that more of the naturally occurring variation in the training and development set is retained in the test set. Local recall compares the content words in the generated output to those in the reference in the test set. A higher score on lexical variation metrics, paired with a high score on local recall, indicates that higher diversity was not obtained at the cost of meaningful content words.

Regarding coverage, we again see that the semi-supervised learning approaches consistently outperform the system that only used the standard training set. The highest coverage scores for WebNLG and Dutch CACAPO were achieved with pseudo-labeling, while data augmentation achieved the highest coverage score for English CACAPO and for E2E. Furthermore, Figure 6 shows that coverage scores are generally increased when the training set size is increased with synthetic data. This effect is most pronounced for pseudo-labeling, whereas an increase in training set size does not necessarily lead to higher coverage scores with data augmentation; while training size increases with data augmentation do seem to have a positive effect on coverage for CACAPO, they seem to decrease coverage scores for WebNLG and E2E.

For local recall, we see that the pseudo-labeling approach achieved the highest scores for Dutch and English CACAPO. The small-scale nature of these domains might stand to benefit more from the extra training data that the pseudo-labeling approach offers to retain commonly occurring content words. For WebNLG and E2E we see that the system trained on the original dataset retains the most content words, while the pseudo-labeling approach generally stays relatively close in terms of local recall scores.

The data augmentation approach shows a more sizable drop-off, which makes sense as the content words were augmented for this semi-supervised learning approach.

Ablation Study. Diversity metrics were also applied to the extension to the original study's design where a T5 model's configuration was initialized without the weights associated with the model and the T5 model that was enriched with both data augmentation as well as pseudo-labeling. These extensions assess the strength of the no extension baseline model, and whether combining the two semi-supervised learning approaches helps to fortify their individual strengths with regard to output quality and diversity.

Table 8 indicates that for all datasets, a model initialized without weights performs worse on most diversity metrics compared with the models that include the pretrained model weights. The model does achieve generally higher scores on average sentence length (for the CACAPO datasets) and standard deviation of sentence length (for all corpora), but paired with the lower number of types, coverage, novelty, and local recall scores, this seems to be an indicator of natural language degeneration (Holtzman et al. 2020). However, for the E2E dataset, this dropoff compared to the other models does not seem as sizeable compared with the other datasets. A reason for this may be that the lack of variety and repetition in this dataset, together with its size, allows for ample opportunity for the model to learn the appropriate model weights.

The effects of joint data augmentation and pseudo-labeling can be found in Figure E.2 and Table 8. Similar to the results of the automated metrics, it seems that this combination performs similar to the best-performing semi-supervised learning approach, albeit a bit worse. For the E2E and WebNLG dataset, this approach does seem to output more types, albeit only a small amount more. For other metrics, the combination of the two approaches achieves higher scores, although the scores are almost similar to those achieved with the pseudo-labeling method. Combining the two approaches therefore does not seem to have a marked impact on output diversity.

Summary. In support of H2, we find that semi-supervised learning increases the output diversity. Data augmentation seems to be the best method to achieve this for English CACAPO and E2E, whereas the pseudo-labeling approach achieves better diversity scores for Dutch CACAPO and WebNLG. The results also indicate that the neural NLG system without an semi-supervised learning approach suffers from catastrophic forgetting: Overfitting on the training set too tightly—which means a relative lack of output diversity as a result. We do not find a marked difference between datasets in terms of the effects of semi-supervised learning in general on output diversity. All datasets seem to benefit relatively equally from semi-supervised learning. Therefore, no clear effect of dataset type (crowdsourcing based vs. naturally occurring) and language (Dutch vs. English) was detected, which is contrary to H4 and H5, respectively.

5.2 Quantitative Human Evaluation

In addition to the automatic evaluations, we also conducted a human evaluation to investigate perceived quality, by measuring fluency, correctness, and grammaticality. The overall scores are summarized in Table 9, and reported per sub-corpus in Table B.3. Inspection of Table 9 reveals a clear pattern: In almost all cases the highest scores are obtained with the pseudo-labeling approach.

This was further examined by conducting a series of linear mixed models, with semi-supervised methods and the datasets/dataset domains as independent variables

Table 9
Mean fluency, correctness, and grammaticality per semi-supervised learning type for each dataset (SDs in parentheses). Different superscripts indicate significant differences between semi-supervised learning approaches. Higher scores mean more positively perceived output.

Dataset	**T**rain Type	**N**	**F**luency	**C**orrectness	**G**rammaticality
CACAPO (en)	No_Ext	43	4.83 (2.05)^a	4.93 (1.86)^a	2.90 (0.98)^a
	Dat_Aug		4.82 (2.05)^a	4.92 (1.89)^a	2.93 (0.98)^a
	Pseu_Lab		**5.67 (1.70)**^{b}	**5.57 (1.70)**^{b}	**3.29 (0.86)**^{b}
CACAPO (nl)	No_Ext	41	4.75 (2.03)^a	5.15 (1.90)^a	2.83 (1.09)^a
	Dat_Aug		4.73 (2.09)^a	5.15 (1.88)^a	2.79 (1.12)^a
	Pseu_Lab		**5.94 (1.52)**^{b}	**5.75 (1.58)**^{b}	**3.49 (0.78)**^{b}
WebNLG	No_Ext	99	4.64 (2.08)^a	3.99 (2.21)^a	2.92 (1.02)^a
	Dat_Aug		4.67 (2.08)^a	4.01 (2.22)^a	2.91 (1.02)^a
	Pseu_Lab		**5.74 (1.66)**^{b}	**5.63 (1.82)**^{b}	**3.43 (0.79)**^{b}
E2E	No_Ext	9	5.08 (1.73)^a	**4.25 (1.86)**^{a}	3.06 (0.88)^a
	Dat_Aug		5.06 (1.70)^a	4.20 (1.83)^a	2.98 (0.91)^a
	Pseu_Lab		**6.26 (1.02)**^{b}	3.86 (1.71)^a	**3.61 (0.65)**^{b}

and fluency, correctness, and grammaticality as the dependent variables. Linear mixed models enable us to control for the systematic variation due to the nested/clustered structure of the data caused by participants having rated multiple texts. Furthermore, linear mixed models are able to deal appropriately with unequal sample sizes (i.e., language and dataset in this experiment). To do so, we added dataset, domain, and participant as a nested random factor in the models.

Following van Miltenburg et al. (2019), we used the `lme4` (Bates et al. 2015) package in `R` to build our linear mixed models using the `lmer` function and estimate p-values for the models, respectively. Datasets and dataset domains were investigated using separate models, and a separate model for each dependent variable was also necessary, meaning that a total of 6 models were created. Furthermore, "no extension" served as the reference level intercept for semi-supervised learning approach, and "E2E" as the reference level intercept for dataset and dataset domain. The `emmeans` package was used for pairwise comparisons when analyzing the differences between semi-supervised learning approaches, datasets, dataset domains, and interaction effects if the linear mixed model showed significant main effects or interaction effects. All models converged and the visual check of model assumptions using the residual plots indicated no signs of violations.

Fluency. The model for fluency was significant (conditional $R^2 = 0.23$, marginal $R^2 = 0.07$, $p < .001$). The model showed a main effect of training type, but not of dataset, and also did not suggest any interaction effects. Table 9 represents the mean scores and training type differences, while the mixed model results are summarized in Table C.1.

Significant effects were further investigated using estimated marginal means with Bonferroni correction, which corroborated these findings. Only the pairwise comparisons for semi-supervised learning approaches showed significance: There was no significant difference in fluency between no extension (M = 4.73, SD = 2.05) and data augmentation (M = 4.74, SD = 2.06), but the fluency scores for the pseudo-labeling approach (M = 5.79, SD = 1.62) were significantly higher than both the data augmentation

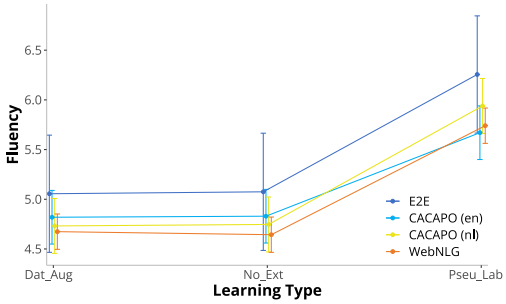

Figure 7
Mean fluency per dataset and learning type.

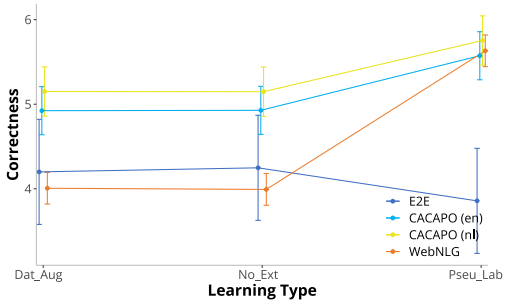

Figure 8
Mean correctness per dataset and learning type.

approach and no extension. This effect was robust across all dataset domains as well, with all but 4 domains showing the same pattern between semi-supervised learning approaches (see Table B.3).

Semi-supervised Learning Type Differences. Figure 7 illustrates these results: All datasets show a similar pattern, where the fluency scores between no extension and data augmentation are virtually similar, and the perceived fluency increases for pseudo-labeling. Sentences were thus perceived as more fluent when a training set was enriched with data created via the pseudo-labeling approach. Furthermore, perceived fluency scores were relatively similar for all the investigated datasets, and the fluency differences between the semi-supervised learning approaches were similar across all datasets.

Summary. Partly in support of H1, semi-supervised learning seems to increase the output quality compared with a language model only trained on the base training set. However, this is only the case for the pseudo-labeling approach and not for the data augmentation approach (RQ1). Furthermore, these fluency results do not support H3, nor H5: The beneficial effect of semi-supervised learning is consistent regardless of whether the original dataset is small-scale or large-scale (H3), Dutch or English (H5).

Correctness. For correctness, the model was significant as well (conditional $R^2 = 0.29$, marginal $R^2 = 0.11$, $p < .001$). The model showed a main effect of semi-supervised learning approach, dataset, and an interaction between the two. Table 9 shows the mean scores and semi-supervised learning approach differences, and Table C.1 a summary of the mixed models for correctness.

The significant effects were further explored using estimated marginal means with Bonferroni correction. For semi-supervised learning type, there was no significant difference between no extension ($M = 4.46$, $SD = 2.12$) and the data augmentation approach ($M = 4.47$, $SD = 2.12$), but there were significantly higher correctness scores for the pseudo-labeling approach ($M = 5.56$, $SD = 1.78$) compared with both other approaches.

For dataset, there was no significant difference between E2E ($M = 4.10$, $SD = 1.80$) and WebNLG ($M = 4.54$, $SD = 2.23$), and no difference between English CACAPO ($M = 5.14$, $SD = 1.84$) and Dutch CACAPO ($M = 5.35$, $SD = 1.81$). However, both CACAPO datasets scored significantly higher on perceived correctness than both WebNLG and E2E. The main effects and interaction effect can be explained in more detail using Figure 8.

Dataset × Semi-supervised Learning Type Differences. Regarding differences between datasets, differentiated per semi-supervised learning approach, we see that for no extension, WebNLG (M = 3.99, SD = 2.21) obtained significantly worse correctness scores compared with both Dutch CACAPO (M = 5.15, SD = 1.90) and English CACAPO (M = 4.93, SD = 1.86).

This pattern is similar when comparing the datasets enriched with data augmentation: WebNLG (M = 4.01, SD = 2.22) performed significantly worse in terms of correctness compared with both Dutch CACAPO (M = 5.15, SD = 1.88) and English CACAPO (M = 4.92, SD = 1.89). This consistency also makes sense when comparing the datasets' scores on no extension to data augmentation: None of the datasets' scores on data augmentation differ significantly from the scores they obtained on no extension.

The pattern is different for pseudo-labeling: The correctness scores for English CACAPO (M = 5.57, SD = 1.70), Dutch CACAPO (M = 5.75, SD = 1.58), and WebNLG (M = 5.63, SD = 1.82) are significantly higher for pseudo-labeling than they are for both no extension and data augmentation, resulting in no significant differences in correctness scores between the three datasets for pseudo-labeling.

However, the correctness score for E2E on pseudo-labeling (M = 3.86, SD = 1.71) is not significantly different compared to the correctness score E2E obtained on no extension and data augmentation. This results in E2E performing significantly worse on pseudo-labeling compared to English CACAPO, Dutch CACAPO, and WebNLG.

Dataset Domain Differences. On a dataset domain level, the `Astronaut`, `Food`, and `Monument` domains of WebNLG in particular performed worse for WebNLG when the base training set or data augmentation was used. Furthermore, we see that the differences between semi-supervised learning approaches are most pronounced for the WebNLG domains, while the differences are not as clear for the English CACAPO and Dutch CACAPO domains (see Table B.3).

Summary. The data shows that overall, no extension and data augmentation performed similarly on perceived correctness across all datasets, and that pseudo-labeling also leads to the highest correctness scores as it does for fluency scores. However, this effect is moderated by the dataset. E2E did not achieve higher correctness scores with the pseudo-labeling approach compared to the other approaches, while the other datasets did.

Similar to the fluency results, the correctness scores are partially in support of H1, showing that the semi-supervised learning shows higher correctness scores compared with a language model only trained on the base training set, but only in the case of the pseudo-labeling approach and not for the data augmentation approach (RQ1). The correctness scores are partially in support of H3, as the beneficial effect of semi-supervised learning could not be found for the largest-scale dataset (E2E), but it could be found for WebNLG. Finally, the correctness scores do not seem to differ per language, which does not support H5.

Grammaticality. The model for grammaticality was significant (conditional R^2 = 0.20, marginal R^2 = 0.07, $p < .001$). Similar to the fluency model, the grammaticality model showed a main effect of training type, but not of dataset, and also did not suggest any interaction effects. Mean scores for grammaticality and differences for training type are shown in Table 9 and the mixed model results are summarized in Table C.1.

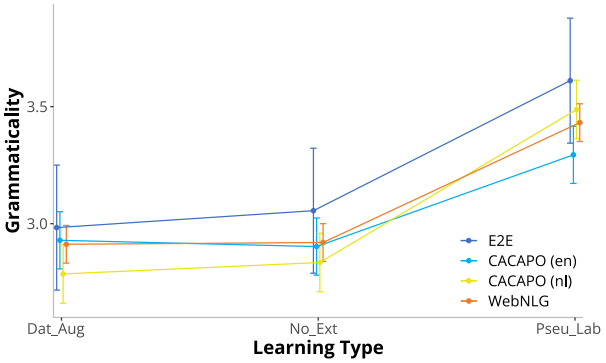

Figure 9
Mean grammaticality per dataset and learning type.

The estimated marginal means with Bonferroni correction shows that no extension (M = 2.90, SD = 1.02) was not significantly different compared with data augmentation (M = 2.89, SD = 1.03) in terms of grammaticality, and that both no extension and data augmentation obtained significantly worse grammaticality scores compared with pseudo-labeling (M = 3.42, SD = 0.80).

Semi-supervised Learning Type Differences. Figure 9 shows a similar pattern as was found for fluency. The pseudo-labeling approach led to higher grammaticality scores than no extension and data augmentation did. Furthermore, the effects of grammaticality seem to be consistent among datasets: The datasets did not differ significantly from each other in terms of the obtained grammaticality scores, and performed similarly with the semi-supervised learning approaches.

Dataset Domain Differences. The differences between semi-supervised learning approaches are also fairly stable on a category level, with 14 out of 19 domains having no significant difference between no extension and data augmentation, and pseudo-labeling performing significantly better than the other two approaches.

Summary. Similar to fluency, the results for grammaticality are partly in support of H1, as semi-supervised learning increased grammaticality scores compared to a language model only trained on the base training set, but only for the the pseudo-labeling approach (RQ1). The grammaticality results also do not support H3/H5 as dataset size (H3) and language (H5) did not alter the found effect of semi-supervised learning.

5.3 Error Analysis

The goal of the error analysis was to investigate whether the different systems were prone to other kinds of errors, which could indicate points of attention for future work. Hence, a chi-square test was conducted with error type (see Section 4.3), semi-supervised learning approach (no extension, data augmentation, and pseudo-labeling), and dataset (Dutch CACAPO, English CACAPO, WebNLG, and E2E) as categorical

Table 10
Number of errors for each dataset per error category and semi-supervised learning approach. Different superscripts indicate significant differences between semi-supervised learning approaches.

CACAPO (NL)	NoExt	DatAug	PseuLab	Total
Hallucinations	6^{a}	7^{a}	5^{a}	18
Missing info	4^{a}	3^{a}	3^{a}	10
References	2^{a}	3^{a}	1^{a}	6
Verb form	4^{a}	2^{a}	1^{a}	7
Determiners	2^{a}	1^{a}	0^{a}	3
Punct./capital.	0^{a}	0^{a}	0^{a}	0
Lexical choices	4^{a}	4^{a}	2^{a}	10
Repetition	3^{a}	3^{a}	1^{a}	7
Connections	7^{a}	4^{a}	1^{a}	12
Missing parts	4^{a}	2^{a}	2^{a}	8
Other	2^{a}	3^{a}	2^{a}	7
Total	38^{a}	$32^{a,b}$	18^{b}	88

CACAPO (EN)	NoExt	DatAug	PseuLab	Total
Hallucinations	9^{a}	10^{a}	5^{a}	24
Missing info	2^{a}	0^{a}	0^{a}	2
References	0^{a}	0^{a}	1^{a}	1
Verb form	2^{a}	2^{a}	1^{a}	5
Determiners	0^{a}	1^{a}	0^{a}	1
Punct./capital.	2^{a}	1^{a}	1^{a}	4
Lexical choices	3^{a}	4^{a}	0^{a}	7
Repetition	2^{a}	5^{a}	0^{a}	7
Connections	2^{a}	4^{a}	1^{a}	7
Missing parts	5^{a}	5^{a}	2^{a}	12
Other	3^{a}	5^{a}	5^{a}	13
Total	$30^{a,b}$	37^{a}	16^{b}	83

WebNLG	NoExt	DatAug	PseuLab	Total
Hallucinations	5^{a}	5^{a}	5^{a}	15
Missing info	11^{a}	12^{a}	6^{a}	29
References	3^{a}	7^{a}	4^{a}	14
Verb form	1^{a}	2^{a}	1^{a}	4
Determiners	1^{a}	2^{a}	1^{a}	4
Punct./capital.	1^{a}	0^{a}	0^{a}	1
Lexical choices	2^{a}	1^{a}	0^{a}	3
Repetition	8^{a}	7^{a}	0^{a}	15
Connections	2^{a}	3^{a}	3^{a}	8
Missing parts	6^{a}	9^{a}	2^{a}	17
Other	$2^{a,b}$	0^{b}	4^{a}	6
Total	$42^{a,b}$	48^{a}	26^{b}	116

E2E	NoExt	DatAug	PseuLab	Total
Hallucinations	9^{a}	10^{a}	13^{a}	32
Missing info	$5^{a,b}$	3^{b}	10^{a}	18
References	8^{a}	7^{a}	4^{a}	19
Verb form	1^{a}	0^{a}	0^{a}	1
Determiners	2^{a}	1^{a}	0^{a}	3
Punct./capital.	2^{a}	4^{a}	0^{a}	6
Lexical choices	1^{a}	0^{a}	0^{a}	1
Repetition	4^{a}	5^{a}	0^{a}	9
Connections	0^{a}	0^{a}	0^{a}	0
Missing parts	1^{a}	2^{a}	0^{a}	3
Other	2^{a}	5^{a}	0^{a}	7
Total	35^{a}	37^{a}	27^{a}	99

All Datasets	NoExt	DatAug	PseuLab	Total
Hallucinations	29^{a}	32^{a}	28^{a}	89
Missing info	22^{a}	18^{a}	19^{a}	59
References	13^{a}	17^{a}	10^{a}	40
Verb form	8^{a}	6^{a}	3^{a}	17
Determiners	5^{a}	5^{a}	1^{a}	11
Punct./capital.	5^{a}	5^{a}	1^{a}	11
Lexical choices	10^{a}	9^{a}	2^{a}	21
Repetition	17^{a}	20^{a}	1^{b}	38
Connections	11^{a}	11^{a}	5^{a}	27
Missing parts	16^{a}	18^{a}	6^{a}	40
Other	9^{a}	13^{a}	11^{a}	33
Total	145^{a}	154^{a}	87^{b}	386

variables. Pairwise comparisons per category between semi-supervised learning types were made using Bonferroni-adjusted z-tests for column proportions. Pairwise comparisons between the total scores per dataset were done using binomial tests with Bonferroni-adjusted *p*-values. Results of the comparisons are shown in Table 10.

Dataset Differences. The chi-square test did not show a significant difference in error proportions between the semi-supervised learning approaches for Dutch CACAPO ($\chi^2(18) = 6.45$, $p = .994$), English CACAPO ($\chi^2(20) = 17.82$, $p = .599$), WebNLG ($\chi^2(20) = 21.11$, $p = .391$), or E2E ($\chi^2(18) = 28.52$, $p = .055$), as well as all the datasets combined ($\chi^2(20) = 26.43$, $p = .152$).

However, the binomial tests with Bonferroni-adjusted *p*-values showed that the proportion of total errors for the pseudo-labeling approach, across all datasets, was significantly lower than the expected (33%, $p < .001$). No difference in proportion of

errors was found between no extension and data augmentation. The pseudo-labeling approach yielded only around half of the errors of the other two approaches in total, and this effect was reasonably consistent for every dataset (see Table 10).

This indicates that differences found in the quantitative human analysis are generally not the result of distinct categorical error differences, but rather an overall difference in all errors combined.

Error Category Differences. Despite no overall effect being found, the pairwise comparisons do show a few significant differences. For WebNLG, a larger proportion of *Other* errors were found for the pseudo-labeling approach (15.4%) compared with the data augmentation approach (0.0%), but this difference is likely explained by the small number of error observations for this particular error category and dataset.

More interesting is the difference found for the E2E dataset regarding the number of missing information errors obtained by the data augmentation approach (8.1%), compared with the pseudo-labeling approach (37%), which could be the reason why the correctness score of the pseudo-labeling approach was lower for the E2E dataset compared with the other datasets in the quantitative human evaluation study.

Furthermore, the errors of all datasets combined showed that the pseudo-labeling approach had much fewer issues with the repetition of words/phrases (1.1% of errors for pseudo-labeling, compared with 11.7% and 13.0% for no extension and data augmentation, respectively). These issues are seen in sentences such as "*Batchoy is eaten in the their, a the of the their spoken is is.*" and may indicate an issue of underfitting, which is solvable by extending the training set via the pseudo-labeling approach. These findings are further elaborated upon with the qualitative evaluation.

5.4 Qualitative Analysis

In addition to the error analysis, the output of the system was further analyzed by focusing on the 15 lowest and 15 highest rated outputs per dataset $\times$ semi-supervised learning approach according to the human evaluation task. The bottom/top 5 of these outputs for each dataset $\times$ semi-supervised learning approach is displayed in Appendix D.

Highest Rated Output. When looking at the highest rated texts, all texts obtained perfect or near-perfect ratings in the human evaluation task. Manual inspection also confirms that none of these sentences contain any clear errors: The texts are grammatical and capture the data (almost always) completely and comprehensively. Furthermore, on a general level it can be observed that the data–text pairs for no extension and data augmentation partially overlap, which further emphasizes the similar performance of both models.

Certain trends could also be deduced from these highest rated texts on a closer inspection. For the Dutch CACAPO dataset, output from the `Weather` and `Accidents` domain accounted for nearly 70% of the highest rated texts. This makes intuitive sense as texts from these domains in the training set are the most consistent in writing style. Furthermore, the data representations for these domains seem to be especially capable of capturing the most important information of the sentence. For instance, the data representation `timePoint="In_the_evening" weatherChange="gradually" weatherType="dry"` becomes "*In the evening it gradually becomes dry*", which is quite close to the data representation.

This trend of a close link between the underlying data and the text leading to the highest scores is also visible for the other datasets. For the English CACAPO dataset, the stocks domain accounted for almost 60% of the highest rated texts. Especially prevalent is a template that contains the `exchangeName`, `stockChange`, `stockPoints`, and `stockChangePercentage` data types (e.g., `exchangeName="Dow" stockChange="up" stockPoints="288.38" stockChangePercentage="1%"` → *"The Dow is up 288.38 points, or 1%."*). This template is likely to appear frequently in the training set with highly regular text representations, which in turn made it easier for the models to learn how to convert this data template into text.

For WebNLG, the models seem to find it easier to convert data representations with fewer triples into text. For no extension and data augmentation, there are no domains that clearly dominate the list of highest rated output. However, for pseudo-labeling, the top 6 are all outputs from the `Building` domain. WebNLG has data representations that stay relatively close to the domain's topic (e.g., for the building domain this would be triples with predicates such as `architect`, `floorCount`, and `location`), and data representations that stray further away from the domain's topic (e.g., for the building domain this would be triples with predicates such as `demonym`, or `leaderName`). The highest rated texts all seem to be the first type of data representations. This especially holds true for pseudo-labeling, which makes sense given that the enriched texts will likely be closer to the domain's topic.

It is difficult to gauge which conditions make the models more effective at producing high quality output for E2E. At least for no extension and data augmentation, similar data representations can be found for the highest and lowest rated texts. For pseudo-labeling, most of the highest rated texts have a similar data representation, with `name`, `eatType`, `food`, `area`, `familyFriendly`, and `near` as the data types (e.g., `name="Giraffe" eatType="restaurant" food="French" area="riverside" familyFriendly="no" near="Raja Indian Cuisine"`. This is likely a frequent data representation template in the training set.

Lowest Rated Output. The lowest rated output in the human evaluation task clearly shows issues in the output. The most common issues seem to be in the production of grammatically correct texts, possibly due to underfitting. Hallucination and omission of data happen somewhat frequently, but making the connections between data points seems to have been the most difficult for the models. Here, overlap between no extension and data augmentation can also be found, in line with the similar performance found between the two systems in the human evaluation.

For the Dutch CACAPO dataset, almost 70% of the lowest rated texts are from the `Weather` and `Sports` domain. For the `Weather` domain, this has mostly to do with texts that verbalize wind information (e.g., *"Tonight the wind will become northeasterly in more and more places."*). For the `Sports` domain, it seems like the variety and lack of repetition in the training set has caused issues with producing grammatically correct texts. For most of these lowest rated texts, the issue is mostly with incorrect use of words, omission, or repetition of words (e.g., *"The man was tripped by an accident."*, *"The south to southwestern is moderate."*). In other cases, mostly the texts related to wind direction, the output is not necessarily incorrect, but may be uncommon information for participants to read.

For the English CACAPO dataset, two thirds of the lowest rated texts are from the `Accidents` and `Stocks` domain. The main issue for this dataset also does not seem to be hallucination or omission of data, but rather issues with the verbalized texts, possibly caused by underfitting (e.g., *"CLEVELAND, Ohio – A November 2014 quintuple*

homicide in Cleveland left one dead and four injured injured injured."). The pseudo labeling approach suffers less from underfitting, but the worst rated texts may be caused by table information in the original articles that have been interpreted as text. For instance "*9pm: -8*" is not a full sentence, but would provide enough information in a table.

Almost half of the lowest rated texts in the WebNLG dataset derive from the `Food` domain, although it is difficult to say why this domain is the most challenging for the models. In all cases, it seems that repetition and incorrect word use are the main issue that cause the low ratings (e.g., "*Amatriciana sauce is a traditional traditional from the Lazio region.*", "*The league is the Greeks of which the team played in the in.*"). For pseudo-labeling, the previously mentioned distinction between close and further removed data representations seems to be the source of difficulties. In most cases for pseudo-labeling, the lowest rated sentences have data representations that are not close to the domain's main topic.

For the lowest rated texts in the E2E dataset, it is also difficult to pinpoint the exact conditions that cause the models to struggle as the data representations seem similar to those of the highest rated texts. However, for pseudo-labeling, it seems that the high frequency of data types in the (enriched) dataset that are not frequent in the test set cause issues. The model hallucinates food types that are not represented in the data (e.g., "*The Cricketers is a restaurant that serves* ***Samoan*** *food.*"). These food types (e.g., *Samoan*, *Burmese*, *Saint Helenian*) are only found in the enriched data for pseudo-labeling, where they occur regularly, which explains why only for E2E, pseudo-labeling had no positive effect on the correctness ratings in the human evaluation study.

6. Discussion and Conclusion

This study investigated the potentially beneficial effect of semi-supervised learning in combination with a language model. More specifically, it investigated whether enriching a training set via the data augmentation approach (i.e., generating several variants of a training text by replacing certain words with synonyms and semantically similar words) and a pseudo-labeling approach (i.e., labeling unlabeled texts using an information extraction model trained on the existing labeled training data) could increase the performance of data-to-text NLG that already utilizes a large-scale language model (T5-large/mT5-large).

Previous work has found that semi-supervised learning could increase the output quality of an NLG system (e.g., Qader, Portet, and Labbé 2019; Schmitt et al. 2020; Su, Huang, and Chen 2020; Tseng et al. 2020; Chang, Demberg, and Marin 2021) and that utilizing language models for data-to-text NLG could help improve output quality as well (e.g., Kale and Rastogi 2020). However, not much is known about the combination of language models and semi-supervised learning in a data-to-text generation setting.

Therefore, it is not known whether semi-supervised approaches are still effective when language models are also used, and, if they are, under what conditions they are effective. Besides the type of semi-supervised learning as a condition, this study also investigated multiple datasets with different characteristics to see whether they affected the effectiveness of semi-supervised learning in combination with language models.

Output Quality. Partial support was found for the hypothesis that semi-supervised learning positively affects output quality (H1). When observing the results of the quantitative human evaluation, the pseudo-labeling approach consistently obtained higher scores on correctness, grammaticality, and fluency compared with the data-to-text

system with a language model that was only trained on the base training set (hereafter: No extension approach).

This is in line with Sun et al. (2020), who found for a text classification task that a combination of a language model and pseudo-labeling led to the highest scores. It also corresponds with previous research that used joint learning systems for data-to-text generation (e.g., Qader, Portet, and Labbé 2019), Schmitt et al. 2020), Su, Huang, and Chen 2020), Tseng et al. 2020), Chang, Demberg, and Marin 2021), with the difference that this research compared it to a system where a language model was finetuned on the training set, rather than a system that was trained on merely the training set.

However, the data augmentation approach performed equivalently to the no extension approach in the quantitative human evaluation study and error analysis, and the output quality scores were generally worse than the quality scores yielded by the pseudo-labeling approach. These results suggest that the pseudo-labeling approach is a better semi-supervised learning approach than the data augmentation approach if the goal is to increase the output quality over a system that is only trained on the base training set (RQ1).

These results make intuitive sense: The data augmentation approach only makes subtle changes to the data that is represented in the text, and does not make any fundamental adjustments to the sentence structure compared with the original text, thus keeping the output relatively similar to that of the no extension approach. The pseudo-labeling approach introduces not only new sentence structures to the training set, but also a large amount of new data types and combinations of data. This increased variety in training data might lead to better generalizations and better handling of the diverse situations in the test set, subsequently leading to higher quality output and less underfitting issues (as was corroborated by the error analysis).

Thus, while previous research with different data augmentation approaches found data augmentation to increase performance compared with a system that was only trained on the training set (e.g., Kulhánek et al. 2021; Riabi et al. 2021; Tandon et al. 2018; Alberti et al. 2019; Chang et al. 2021; Kedzie and McKeown 2019), it may not result in text quality improvements if a language model is also implemented. It is possible that when the original training set is too small (which is common for NLG datasets), it may lead to an underfitted system. Underfitting can be reduced by implementing a language model that is finetuned on the training data, and by extending the training data with data augmentation. However, this combination of the two might be redundant for handling the underfitting issue.

The automatic metrics for output quality estimation suggest that the pseudo-labeling approach is the most effective approach for the CACAPO dataset—especially the Dutch CACAPO dataset—and that the no extension approach leads to the best output for E2E, and WebNLG (in concordance with H3 which states that smaller-scale datasets benefit more from semi-supervised learning, and H5 that states that Dutch datasets benefit more than English datasets).

However, the results of the quantitative human evaluation shows that the fluency and grammaticality outcomes are fairly consistent among datasets. The exception being correctness, where semi-supervised learning did not return higher ratings than the no extension approach for E2E. It is possible that this has more to do with the nature of the extensions used with E2E for the pseudo-labeling approach, as there was a clear correctness increase for the pseudo-labeling approach over the no extension approach for WebNLG.

The semantic parser that was applied on the extra texts for the pseudo-labeling approach seemed to be struggling (with F1 scores of only around 57% on the additional

data), possibly due to overfitting on the original training set. These issues with the labels then eventually affected the correctness scores of the data-to-text system's output (an example of cascading of errors [Castro Ferreira et al. 2019]). In turn, this would lead to worse connections between the data input and text output, as is also found by the error analysis.

Therefore, we believe our results are not in line with H3, nor H5 as the Dutch CACAPO corpus performed similarly to the other corpora, especially the English CACAPO corpus. These results are also not in line with previous research that found semi-supervised learning to especially bolster the results for small-scale datasets (Chang, Demberg, and Marin 2021). It might therefore be that the usage of language models nullifies the differences between datasets and their scale: Previous work has shown that language models enable few-shot and zero-shot learning (Bender et al. 2021), which could make the difference between dataset scale less important. The lack of difference between the Dutch and English datasets furthermore shows that, although the Dutch representation in mT5 is small (especially compared with the English representation in T5) it is still enough to overcome issues of underfitting.

Output Diversity. The diversity differences between the various approaches were investigated using measures derived from van Miltenburg, Elliott, and Vossen (2018). We found support for H2: Semi-supervised learning generally seemed to increase output diversity. For every investigated dataset, one of the semi-supervised learning approaches scored the highest on almost each of the diversity metrics. This is also in line with previous findings by Kulhánek et al. (2021). This suggests that semi-supervised learning can (at least partially) solve the previously highlighted "catastrophic forgetting" problem (Greco et al. 2019), where a neural model is overfit too tightly during finetuning, which leads to the model forgetting about the diverse language in the language model.

The increase in diversity is relatively consistent among datasets (which does not support H4 that states that a larger increase in diversity is expected for crowdsourced datasets) and dataset language (which does not support H5). However, the semi-supervised learning approach that leads to the largest diversity increases differs per dataset. The pseudo-labeling approach is generally the most diversity-bolstering approach for CACAPO (nl) and WebNLG, whereas the data augmentation approach is the most effective approach for diversity increases for CACAPO (en) and E2E.

However, it is difficult to pinpoint which characteristics of the datasets lead to these differences. Diversity differences may be due to the richness in content words in the original dataset (which leads to more perturbations with the data augmentation approach), or lack thereof. Alternatively, it can be that the extra texts used for the pseudo-labeling approach contain more (or less) diverse language for some datasets, leading to more (or less) diverse verbalizations of the data-to-text output. Future research is necessary to investigate which characteristics of the dataset—and extra data introduced by the semi-supervised learning approaches—are most salient when the goal is to increase text diversity.

6.1 Future Work

The current study found that semi-supervised learning is an effective technique for data-to-text generation, even when used in conjunction with a language model. The pseudo-labeling approach can increase both output diversity and quality, whereas the data augmentation approach is effective at increasing diversity, while keeping the quality consistent with a model only enriched with a language model.

Furthermore, this result seems to be consistent among datasets. It may be that a good language model nullifies the differences between the original datasets, meaning that, for instance, the scale of the dataset does not matter as much. This in turn would mean that enriching training data with a language model can already be helpful for non-English NLG tasks, for which it is generally more difficult to find large-scale datasets (Riabi et al. 2021). This would help to make neural NLG systems available for more people around the world without having to invest in large-scale datasets.

Two common semi-supervised learning approaches (pseudo-labeling and data augmentation) were explored in the design of the presented data-to-text system. However, for future research there are also other approaches to explore. For instance, data augmentation can be done in other ways, other than perturbing words with semantically similar words, as was done in this research, such as data augmentation in the form of self-training where new texts are generated from unlabeled data. Previous research has shown that such an approach to data augmentation is promising (e.g., Heidari et al. 2021; Jolly et al. 2022; Mehta et al. 2022).

Also of interest would be *entity resolution/text matching* (i.e., automatically connecting data with corresponding text), rather than pseudo-labeling (Ahmadi, Sand, and Papotti 2021). Many real-world companies have large collections of texts and related data, but are missing an explicit connection between the two (van der Lee et al. 2020). Such a task could help to produce large quantities of extra data, and may additionally help to tackle issues with mistakes in the original manual data annotations that could lead to bad texts.

Especially interesting would be the use of unlabeled texts as an in-domain language model, rather than the Transformers-based language models such as BERT and T5 that were trained on an immense variety of different domains. Sun et al. (2020) found that using such an in-domain language model made the use of BERT unnecessary, and that the combination of in-domain pretraining and pseudo-labeling resulted in the best model for a text classification task. The impact of training large Transformers models such as BERT and T5 on CO_2 emissions is well documented (Strubell, Ganesh, and McCallum 2019). If the use of unlabeled texts for (generally smaller-scale) in-domain pretraining and pseudo-labeling is indeed also effective for NLG, this might help to focus development on more efficient and eco-friendly development of single-domain language models, rather than unsustainably striving for larger and larger all-encompassing Transformers models.

The findings of the current study highlight the feasibility of automatically extending datasets in a way that neural data-to-text systems benefit from the extension, even if the original dataset is small in size. This can be beneficial for robot journalism, for instance, which still mostly relies on rule-based NLG. Newspaper organizations oftentimes have large archives of texts and data, but aligning the two manually on a scale that permits a neural model to be trained on it requires an amount of manpower that most newsrooms do not have available. By creating a small starting point and then using semi-supervised learning to create the full dataset, would be a less intensive undertaking—thus a smaller hurdle—for many companies to use neural data-to-text systems.

Appendix A. Examples of Output

Table A.1
Random examples of output for each of the datasets and semi-supervised learning methods.

Dataset	Train type	Input data	Output text
CACAPO (en)	No_Ext	victimAgeGroup="adult" victimSex="female" victimStatus="dead"	Police say an adult female was found dead at the scene from apparent gunshot wounds.
	Dat_Aug	batterName="Xavier_Nady" pitchResult="scored" teamName="Cubs" inningNumber="sixth"	Xavier Nady scored for the Cubs in the sixth.
	Pseu_Lab	companyName="Microsoft" stockChange="jumped" stockChangePercentage="more_than_ 54%" moneyAmount1="$138.40" moneyAmount2="$90.14" timePoint="Jan._17"	Microsoft shares jumped more than 54 % to $138.40, from $90.14 on Jan. 17.
CACAPO (nl)	No_Ext	incidentLocation="A12_richting_Arnhem" dateTime="dinsdag" incidentType="dodelijk_ongeluk"	De A12 richting Arnhem is dinsdag door een dodelijk ongeluk geindigd. *The A12 to Arnhem has ended by a deadly accident on Tuesday.*
	Dat_Aug	hasWonTeam="AZ" matchDate="donderdagavond" finalScore="1-2" hasLostTeam="ADO_Den_Haag"	AZ heeft donderdagavond met 1-2 gewonnen van ADO Den Haag. *AZ has won 1-2 against ADO Den Haag on Thursday night.*
	Pseu_Lab	weatherType="regen" compassDirection="noordoosten"	De regen trekt verder naar het noordoosten. *The rain will move further to the northeast.*
WebNLG	No_Ext	"Turkey" — capital — "Ankara" "Turkey" — leaderName — "Ahmet Davutolu"	The capital of Turkey is Ankara and the country's leader is Ahmet Davutoglu.
	Dat_Aug	"Tim Brooke-Taylor" — birthPlace — "Derbyshire" "Bananaman" — starring — "Tim Brooke-Taylor"	Derbyshire born Tim Brooke Taylor starred in The film.
	Pseu_Lab	"Buzz Aldrin" — occupation — "Fighter pilot" "Buzz Aldrin" — was_a_crew_member_of — "Apollo 11"	He was a fighter pilot and a member of the Apollo 11 crew.
E2E	No_Ext	name="The Punter" eatType="pub" food="English" priceRange="high" area="riverside" familyFriendly="no" near="Raja Indian Cuisine"	The Punter is a high priced pub near Raja Indian Cuisine in the riverside area. It is not children friendly.
	Dat_Aug	name="Clowns" eatType="pub" customer rating="3 out of 5" near="All Bar One"	Clowns pub is a local - priced, local - rated, and - friendly - friendly coffee bar located in the city market, near All Bar One.
	Pseu_Lab	name="The Cricketers" eatType="restaurant" customer rating="low" familyFriendly="no" near="Ranch"	The Cricketers is a restaurant located near Rancho Mexican Cafe. It is not family-friendly.

Appendix B. Evaluation Results per Dataset Domain

Table B.1
Automatic metric results of the different (XL-format) semi-supervised learning approaches (No_Ext = no training set extension, Dat_Aug = data augmentation, Pseu_Lab = pseudo-labeling) for each dataset domain (bold = highest).

Dataset	Domain	Train type	BLEU	NIST	BERTScore	METEOR	ROUGE-L
CACAPO	Incidents (en)	No_Ext	29.47	5.11	58.55	51.28	47.08
		Dat_Aug	28.15	4.99	53.71	47.18	42.92
		Pseu_Lab	**34.51**	**5.52**	**62.06**	**54.75**	**52.07**
CACAPO	Sports (en)	No_Ext	31.77	6.63	61.00	56.43	52.73
		Dat_Aug	23.30	5.80	52.57	48.76	46.35
		Pseu_Lab	**37.28**	**7.23**	**65.55**	**60.02**	**57.30**
CACAPO	Stocks (en)	No_Ext	25.95	4.94	50.03	50.98	45.62
		Dat_Aug	23.02	5.14	44.50	48.23	42.55
		Pseu_Lab	**32.44**	**6.05**	**54.24**	**55.08**	**50.72**
CACAPO	Weather (en)	No_Ext	33.95	6.24	65.88	62.61	56.36
		Dat_Aug	25.80	5.29	57.27	50.29	49.71
		Pseu_Lab	**39.59**	**6.70**	**70.34**	**66.99**	**62.97**
CACAPO	Incidents (nl)	No_Ext	34.22	5.39	85.15	53.38	51.53
		Dat_Aug	36.71	5.01	87.20	55.50	57.10
		Pseu_Lab	**43.39**	**6.54**	**87.65**	**60.20**	**60.10**
CACAPO	Sports (nl)	No_Ext	18.05	4.40	81.75	41.35	42.27
		Dat_Aug	18.45	4.35	81.75	42.13	43.04
		Pseu_Lab	**25.45**	**5.19**	**83.42**	**48.11**	**48.32**
CACAPO	Stocks (nl)	No_Ext	51.69	7.84	88.51	67.00	64.75
		Dat_Aug	45.04	7.40	87.63	64.91	62.33
		Pseu_Lab	**63.31**	**8.98**	**91.03**	**74.83**	**72.75**
CACAPO	Weather (nl)	No_Ext	24.28	4.61	82.50	47.29	46.44
		Dat_Aug	49.07	7.32	90.42	69.39	67.75
		Pseu_Lab	**76.66**	**10.07**	**95.37**	**84.84**	**84.36**
WebNLG	Airport	No_Ext	**51.54**	**7.35**	**74.63**	**74.80**	**61.17**
		Dat_Aug	25.94	4.66	48.51	50.72	42.62
		Pseu_Lab	50.61	7.16	72.29	72.75	60.46
WebNLG	Astronaut	No_Ext	**48.89**	**6.54**	**77.98**	**76.03**	**65.04**
		Dat_Aug	21.04	3.82	55.69	47.77	43.08
		Pseu_Lab	45.48	6.27	73.41	72.45	60.98
WebNLG	Building	No_Ext	**53.48**	**7.84**	**76.93**	**78.06**	**65.97**
		Dat_Aug	34.51	5.77	60.05	62.70	52.65
		Pseu_Lab	50.46	7.59	73.84	74.87	63.64
WebNLG	City	No_Ext	**29.01**	2.50	**52.93**	53.23	**47.14**
		Dat_Aug	21.09	2.42	47.79	47.46	41.03
		Pseu_Lab	27.83	**4.18**	52.00	**54.64**	44.84
WebNLG	ComicsCharacter	No_Ext	**48.95**	**6.31**	**72.50**	76.11	58.69
		Dat_Aug	38.06	5.56	61.60	69.43	51.17
		Pseu_Lab	48.74	6.22	70.92	**76.68**	**62.42**
WebNLG	Food	No_Ext	**46.81**	**7.30**	**71.35**	**70.53**	**56.11**
		Dat_Aug	23.84	4.57	46.37	47.74	40.19
		Pseu_Lab	38.10	6.28	62.91	61.87	48.89
WebNLG	Monument	No_Ext	43.97	5.95	69.70	71.89	56.88
		Dat_Aug	27.88	4.39	53.21	54.12	46.25
		Pseu_Lab	**45.38**	**6.10**	**71.26**	**73.17**	**58.58**
WebNLG	SportsTeam	No_Ext	**46.07**	**6.89**	**72.41**	**72.59**	**59.58**
		Dat_Aug	25.89	4.79	52.68	52.20	42.65
		Pseu_Lab	43.88	6.77	69.53	70.94	58.18
WebNLG	University	No_Ext	**60.07**	**7.34**	**78.10**	**78.12**	**68.93**
		Dat_Aug	29.60	4.59	52.81	53.72	48.12
		Pseu_Lab	55.34	6.89	74.66	75.65	64.27
WebNLG	WrittenWork	No_Ext	**54.39**	**7.45**	**75.86**	**77.09**	**65.23**
		Dat_Aug	36.00	5.60	62.28	61.89	53.28
		Pseu_Lab	52.25	7.16	73.16	74.57	62.48
E2E		No_Ext	**66.05**	**7.08**	**79.40**	**80.21**	**44.97**
		Dat_Aug	28.41	4.15	56.41	62.49	33.40
		Pseu_Lab	50.51	4.65	63.12	60.39	38.48

Table B.2
Average sentence length, standard deviation of sentence length, mean-segmented type-token ratio (TTR), bigram TTR, percentage novel descriptions, coverage, novelty, and local recall with importance class 1 (bold = highest).

Dataset	Domain	Train type	ASL	SDSL	Types	TTR_1	TTR_2	%Novel	Cov	Nov	Loc_1
		No_Ext	16.05	6.76	656	0.63	0.90	99.04	0.53	0.19	0.49
CACAPO	Incidents (en)	Dat_Aug	**17.83**	7.46	**844**	**0.68**	**0.94**	**100.00**	**0.62**	**0.31**	0.47
		Pseu_Lab	15.86	**8.47**	696	0.63	0.89	99.04	0.57	0.20	**0.52**
		No_Ext	19.32	7.95	1,499	0.69	**0.96**	99.84	0.58	0.14	0.57
CACAPO	Sports (en)	Dat_Aug	19.07	8.17	**1,574**	**0.70**	**0.96**	**100.00**	0.59	**0.16**	0.54
		Pseu_Lab	**19.40**	**8.52**	1,559	**0.70**	**0.96**	**100.00**	**0.61**	0.13	**0.61**
		No_Ext	18.15	9.51	1,369	0.65	0.91	99.07	0.49	0.31	0.43
CACAPO	Stocks (en)	Dat_Aug	19.10	9.32	**1,581**	**0.68**	**0.94**	**100.00**	**0.54**	**0.38**	0.45
		Pseu_Lab	**19.76**	**11.10**	1,556	0.65	0.91	99.69	0.53	**0.38**	**0.48**
		No_Ext	**13.62**	7.05	746	0.62	0.91	94.28	0.57	0.14	0.59
CACAPO	Weather (en)	Dat_Aug	12.75	**7.22**	757	**0.65**	**0.92**	**99.18**	0.57	**0.16**	0.57
		Pseu_Lab	13.40	7.10	**768**	0.63	0.91	94.55	**0.60**	0.14	**0.63**
		No_Ext	13.84	5.24	539	0.58	0.86	**100.00**	0.53	0.21	0.52
CACAPO	Incidents (nl)	Dat_Aug	12.34	4.94	526	0.58	**0.87**	**100.00**	0.51	0.22	0.54
		Pseu_Lab	**14.60**	**6.51**	**584**	**0.59**	**0.87**	99.01	**0.58**	**0.23**	**0.58**
		No_Ext	13.78	**5.82**	920	0.66	0.93	**100.00**	0.48	0.10	0.42
CACAPO	Sports (nl)	Dat_Aug	13.43	5.50	990	0.67	0.94	**100.00**	0.49	0.13	0.43
		Pseu_Lab	**13.90**	5.69	**1,127**	**0.70**	**0.96**	99.72	**0.56**	**0.14**	**0.48**
		No_Ext	15.35	7.18	1,373	0.65	0.91	97.54	0.58	0.31	0.62
CACAPO	Stocks (nl)	Dat_Aug	15.17	7.22	1,325	0.65	0.92	**99.11**	0.54	0.32	0.60
		Pseu_Lab	**15.78**	**7.26**	**1,584**	**0.67**	**0.93**	96.20	**0.65**	**0.37**	**0.70**
		No_Ext	15.07	**10.27**	289	0.42	0.70	**98.75**	0.52	0.06	0.51
CACAPO	Weather (nl)	Dat_Aug	15.14	5.35	**459**	0.59	**0.88**	97.24	0.76	**0.17**	0.73
		Pseu_Lab	**15.29**	5.31	446	**0.60**	**0.88**	82.21	**0.84**	0.06	**0.86**
		No_Ext	16.77	6.21	425	0.43	0.70	78.17	0.65	0.03	**0.74**
WebNLG	Airport	Dat_Aug	**16.89**	**8.23**	417	0.42	0.70	**99.30**	0.50	**0.17**	0.48
		Pseu_Lab	16.78	6.19	**474**	**0.45**	**0.71**	79.23	**0.68**	0.08	0.71
		No_Ext	17.56	7.42	208	0.47	0.69	74.68	0.53	0.02	**0.71**
WebNLG	Astronaut	Dat_Aug	16.66	7.26	231	0.45	0.69	**98.70**	0.40	**0.21**	0.36
		Pseu_Lab	**17.79**	**7.65**	**259**	**0.51**	**0.77**	76.62	**0.57**	0.12	0.68
		No_Ext	17.60	6.61	458	0.44	0.70	78.26	**0.72**	0.02	**0.79**
WebNLG	Building	Dat_Aug	**17.72**	**6.93**	456	0.43	0.69	**98.81**	0.63	**0.12**	0.60
		Pseu_Lab	17.29	6.52	**473**	**0.46**	**0.73**	76.68	**0.72**	0.05	0.76
		No_Ext	11.24	2.79	182	0.32	0.52	79.35	0.69	0.11	0.51
WebNLG	City	Dat_Aug	11.69	3.14	208	0.37	0.59	**92.90**	0.69	0.23	0.47
		Pseu_Lab	**13.79**	**4.66**	**270**	**0.38**	**0.64**	87.10	**0.81**	**0.37**	**0.55**
		No_Ext	14.97	**5.69**	174	0.45	0.73	95.16	0.64	0.03	**0.70**
WebNLG	ComicsChar	Dat_Aug	14.97	5.02	172	0.45	0.71	**100.00**	0.60	**0.06**	0.60
		Pseu_Lab	**15.39**	5.50	**188**	**0.49**	**0.78**	95.16	**0.66**	**0.06**	0.68
		No_Ext	15.74	**7.53**	439	0.42	0.71	84.10	0.67	0.01	**0.68**
WebNLG	Food	Dat_Aug	**15.91**	7.36	**585**	**0.45**	**0.78**	**98.97**	0.66	**0.24**	0.41
		Pseu_Lab	15.61	7.22	517	0.44	0.72	88.97	**0.71**	0.09	0.59
		No_Ext	19.05	7.35	145	**0.44**	0.71	73.86	0.62	0.04	0.74
WebNLG	Monument	Dat_Aug	18.49	**7.74**	**176**	**0.44**	**0.72**	**98.86**	0.55	**0.26**	0.51
		Pseu_Lab	**19.58**	7.59	152	0.42	0.69	72.73	**0.65**	0.04	**0.75**
		No_Ext	**15.60**	5.46	352	0.47	0.72	88.56	0.64	0.02	**0.66**
WebNLG	SportsTeam	Dat_Aug	15.46	**5.69**	362	0.44	0.71	**100.00**	0.57	**0.11**	0.46
		Pseu_Lab	15.31	5.53	**383**	**0.48**	**0.75**	82.59	**0.68**	0.04	0.63
		No_Ext	16.47	6.78	188	0.43	0.69	67.35	0.60	0.02	**0.75**
WebNLG	University	Dat_Aug	14.86	6.53	227	0.46	0.76	**95.24**	0.53	**0.22**	0.47
		Pseu_Lab	**17.30**	**7.32**	**255**	**0.48**	**0.77**	65.31	**0.73**	0.12	0.72
		No_Ext	18.15	6.78	352	0.41	0.65	84.21	0.61	0.01	**0.74**
WebNLG	Writ.Work	Dat_Aug	17.74	6.49	397	0.44	**0.71**	**97.98**	0.58	**0.12**	0.59
		Pseu_Lab	**18.32**	**6.97**	**400**	**0.45**	**0.71**	82.59	**0.67**	0.04	0.71
		No_Ext	28.58	**7.66**	120	0.34	0.50	**100.00**	0.11	0.00	**0.11**
E2E		Dat_Aug	**34.42**	7.73	**223**	**0.38**	**0.55**	**100.00**	**0.16**	**0.03**	0.10
		Pseu_Lab	23.22	5.26	115	0.26	0.38	**100.00**	0.07	**0.03**	0.08

Table B.3
Mean fluency, correctness, and grammaticality per semi-supervised learning type for each domain (SDs in parentheses). Different superscripts indicate significant differences between semi-supervised learning methods for that domain. Higher scores mean more positively perceived output.

Dataset	Domain	Train Type	N	Fluency	Correctness	Grammaticality
CACAPO	Incidents (en)	No_Ext	11	4.90 (2.16)^{a}	4.77 (1.97)^{a}	2.93 (0.95)^{a}
		Dat_Aug		4.77 (2.18)^{a}	4.77 (2.00)^{a}	2.86 (1.00)^{a}
		Pseu_Lab		**5.99 (1.57)**^{b}	**5.75 (1.82)**^{b}	**3.41 (0.77)**^{b}
CACAPO	Sports (en)	No_Ext	12	4.51 (2.03)^{a}	4.56 (1.80)^{a}	2.83 (0.91)^{a}
		Dat_Aug		4.59 (2.05)^{a}	4.60 (1.92)^{a,b}	2.91 (0.91)^{a}
		Pseu_Lab		**5.42 (1.88)**^{b}	**5.23 (1.82)**^{b}	**3.24 (0.84)**^{b}
CACAPO	Stocks (en)	No_Ext	10	4.95 (2.09)^{a}	5.23 (1.79)^{a}	3.06 (1.07)^{a}
		Dat_Aug		4.80 (2.10)^{a}	5.12 (1.79)^{a}	3.02 (1.02)^{a}
		Pseu_Lab		**5.80 (1.58)**^{b}	**5.82 (1.46)**^{a}	**3.42 (0.78)**^{b}
CACAPO	Weather (en)	No_Ext	10	5.02 (1.85)^{a}	5.24 (1.78)^{a}	2.81 (0.99)^{a}
		Dat_Aug		5.17 (1.81)^{a}	5.29 (1.76)^{a}	2.94 (0.98)^{a}
		Pseu_Lab		**5.48 (1.67)**^{a}	**5.54 (1.57)**^{a}	**3.10 (0.99)**^{a}
CACAPO	Incidents (nl)	No_Ext	11	4.87 (2.05)^{a}	5.23 (1.85)^{a}	2.84 (1.10)^{a}
		Dat_Aug		4.87 (2.14)^{a}	5.25 (1.86)^{a}	2.84 (1.11)^{a}
		Pseu_Lab		**6.29 (1.20)**^{b}	**5.91 (1.40)**^{a}	**3.61 (0.66)**^{b}
CACAPO	Sports (nl)	No_Ext	10	4.73 (2.12)^{a}	5.03 (1.98)^{a}	2.90 (1.10)^{a}
		Dat_Aug		4.75 (2.08)^{a}	4.98 (1.92)^{a}	2.90 (1.10)^{a}
		Pseu_Lab		**5.44 (1.75)**^{a}	**5.34 (1.77)**^{a}	**3.22 (0.96)**^{a}
CACAPO	Stocks (nl)	No_Ext	11	5.00 (1.98)^{a}	5.32 (1.77)^{a}	3.08 (0.97)^{a}
		Dat_Aug		4.95 (2.04)^{a}	5.28 (1.82)^{a}	2.95 (1.08)^{a}
		Pseu_Lab		**5.91 (1.57)**^{b}	**5.67 (1.64)**^{a}	**3.59 (0.69)**^{b}
CACAPO	Weather (nl)	No_Ext	9	4.31 (1.90)^{a}	4.97 (2.03)^{a}	2.45 (1.10)^{a}
		Dat_Aug		4.27 (2.02)^{a}	5.06 (1.92)^{a}	2.38 (1.13)^{a}
		Pseu_Lab		**6.10 (1.41)**^{b}	**6.12 (1.40)**^{b}	**3.52 (0.75)**^{b}
WebNLG	Airport	No_Ext	11	4.23 (2.06)^{a}	4.00 (2.12)^{a}	2.85 (1.02)^{a}
		Dat_Aug		4.23 (2.08)^{a}	3.99 (2.19)^{a}	2.84 (1.00)^{a}
		Pseu_Lab		**5.44 (1.71)**^{b}	**5.40 (1.82)**^{b}	**3.39 (0.83)**^{b}
WebNLG	Astronaut	No_Ext	12	4.93 (2.01)^{a}	3.13 (1.86)^{a}	3.06 (0.96)^{a}
		Dat_Aug		4.89 (1.98)^{a}	3.02 (1.85)^{a}	2.99 (1.00)^{a}
		Pseu_Lab		**6.26 (1.17)**^{b}	**6.17 (1.37)**^{b}	**3.70 (0.58)**^{b}
WebNLG	Building	No_Ext	6	5.23 (1.57)^{a}	4.84 (1.72)^{a}	3.07 (0.82)^{a}
		Dat_Aug		5.26 (1.61)^{a}	4.79 (1.83)^{a}	3.08 (0.76)^{a}
		Pseu_Lab		**6.03 (1.34)**^{a}	**6.14 (1.33)**^{b}	**3.52 (0.69)**^{a}
WebNLG	City	No_Ext	9	**5.42 (1.88)**^{a}	4.81 (2.08)^{a}	**3.27 (1.00)**^{a}
		Dat_Aug		5.40 (1.95)^{a}	**4.84 (2.02)**^{a}	**3.27 (0.99)**^{a}
		Pseu_Lab		5.18 (2.07)^{a}	4.44 (2.27)^{a}	3.20 (0.96)^{a}
WebNLG	ComicsChar	No_Ext	12	5.66 (1.75)^{a}	5.46 (1.78)^{a}	3.37 (0.81)^{a}
		Dat_Aug		**5.73 (1.70)**^{a}	5.48 (1.80)^{a}	**3.39 (0.83)**^{a}
		Pseu_Lab		5.72 (1.76)^{a}	**5.59 (1.84)**^{a}	3.38 (0.84)^{a}
WebNLG	Food	No_Ext	8	3.35 (2.28)^{a}	2.92 (2.18)^{a}	2.15 (1.13)^{a}
		Dat_Aug		3.35 (2.29)^{a}	2.97 (2.09)^{a}	2.14 (1.14)^{a}
		Pseu_Lab		**5.89 (1.69)**^{b}	**5.46 (1.96)**^{b}	**3.41 (0.91)**^{b}
WebNLG	Monument	No_Ext	12	3.99 (2.13)^{a}	2.80 (2.10)^{a}	2.50 (0.97)^{a}
		Dat_Aug		4.13 (2.17)^{a}	2.99 (2.19)^{a}	2.51 (1.02)^{a}
		Pseu_Lab		**5.92 (1.44)**^{b}	**5.83 (1.75)**^{b}	**3.42 (0.72)**^{b}
WebNLG	SportsTeam	No_Ext	11	4.89 (1.89)^{a}	4.64 (2.16)^{a}	3.06 (0.97)^{a}
		Dat_Aug		4.90 (1.87)^{a}	4.66 (2.16)^{a}	3.04 (0.92)^{a}
		Pseu_Lab		**5.96 (1.38)**^{b}	**5.94 (1.54)**^{b}	**3.51 (0.69)**^{b}
WebNLG	University	No_Ext	9	4.60 (1.94)^{a}	3.71 (2.03)^{a}	3.09 (0.85)^{a}
		Dat_Aug		4.70 (1.82)^{a}	3.60 (2.02)^{a}	3.07 (0.79)^{a}
		Pseu_Lab		**5.70 (1.72)**^{b}	**5.60 (1.84)**^{b}	**3.49 (0.73)**^{b}
WebNLG	Writ.Work	No_Ext	9	4.00 (2.05)^{a}	3.82 (2.16)^{a}	2.68 (1.02)^{a}
		Dat_Aug		4.00 (2.07)^{a}	3.91 (2.22)^{a}	2.71 (1.02)^{a}
		Pseu_Lab		**5.21 (1.92)**^{b}	**5.66 (1.78)**^{b}	**3.23 (0.88)**^{b}
E2E		No_Ext	9	5.08 (1.73)^{a}	**4.25 (1.86)**^{a}	3.06 (0.88)^{a}
		Dat_Aug		5.06 (1.70)^{a}	4.20 (1.83)^{a}	2.98 (0.91)^{a}
		Pseu_Lab		**6.26 (1.02)**^{b}	3.86 (1.71)^{a}	**3.61 (0.65)**^{b}

Appendix C. Multiple Mixed Model Linear Regressions

Table C.1
Multiple mixed model linear regressions of fluency, correctness, and grammaticality, respectively.

Parameter	*B*	SE	95% CI	*t*	*p*
(Intercept)	5.08	0.30	[4.49, 5.66]	16.87	< 0.001
Train Type [Dat_Aug]	−0.02	0.18	[−0.38, 0.34]	−0.11	0.916
Train Type [Pseu_Lab]	1.18	0.18	[0.82, 1.54]	6.43	< 0.001
Dataset [CACAPO (en)]	−0.25	0.33	[−0.89, 0.40]	−0.74	0.458
Dataset [CACAPO (nl)]	−0.33	0.33	[−0.98, 0.32]	−0.99	0.323
Dataset [WebNLG]	−0.43	0.31	[−1.05, 0.18]	−1.37	0.170
Train Type [Dat_Aug] × Dataset [CACAPO (en)]	0.01	0.20	[−0.39, 0.40]	0.04	0.965
Train Type [Pseu_Lab] × Dataset [CACAPO (en)]	−0.34	0.20	[−0.74, 0.06]	−1.69	0.092
Train Type [Dat_Aug] × Dataset [CACAPO (nl)]	0.00	0.20	[−0.39, 0.40]	0.02	0.982
Train Type [Pseu_Lab] × Dataset [CACAPO (nl)]	0.01	0.20	[−0.39, 0.41]	0.06	0.953
Train Type [Dat_Aug] × Dataset [WebNLG]	0.05	0.19	[−0.33, 0.43]	0.26	0.794
Train Type [Pseu_Lab] × Dataset [WebNLG]	−0.08	0.19	[−0.46, 0.29]	−0.44	0.661
Parameter	***B***	**SE**	**95% CI**	***t***	***p***
(Intercept)	4.25	0.32	[3.63, 4.87]	13.41	< 0.001
Train Type [Dat_Aug]	−0.05	0.19	[−0.41, 0.32]	−0.26	0.795
Train Type [Pseu_Lab]	−0.39	0.19	[−0.75, −0.03]	−2.11	0.035
Dataset [CACAPO (en)]	0.68	0.35	[0.00, 1.36]	1.95	0.051
Dataset [CACAPO (nl)]	0.90	0.35	[0.21, 1.59]	2.57	0.010
Dataset [WebNLG]	−0.26	0.33	[−0.90, 0.39]	−0.77	0.440
Train Type [Dat_Aug] × Dataset [CACAPO (en)]	0.04	0.20	[−0.36, 0.44]	0.22	0.830
Train Type [Pseu_Lab] × Dataset [CACAPO (en)]	1.04	0.20	[0.64, 1.44]	5.08	< 0.001
Train Type [Dat_Aug] × Dataset [CACAPO (nl)]	0.05	0.20	[−0.35, 0.45]	0.24	0.806
Train Type [Pseu_Lab] × Dataset [CACAPO (nl)]	1.00	0.20	[0.59, 1.40]	4.86	< 0.001
Train Type [Dat_Aug] × Dataset [WebNLG]	0.06	0.19	[−0.32, 0.44]	0.32	0.749
Train Type [Pseu_Lab] × Dataset [WebNLG]	2.03	0.19	[1.65, 2.41]	10.48	< 0.001
Parameter	***B***	**SE**	**95% CI**	***t***	***p***
(Intercept)	3.06	0.14	[2.79, 3.32]	22.43	< 0.001
Train Type [Dat_Aug]	−0.07	0.09	[−0.26, 0.11]	−0.77	0.439
Train Type [Pseu_Lab]	0.56	0.09	[0.37, 0.74]	5.95	< 0.001
Dataset [CACAPO (en)]	−0.15	0.15	[−0.45, 0.14]	−1.02	0.306
Dataset [CACAPO (nl)]	−0.22	0.15	[−0.52, 0.07]	−1.47	0.141
Dataset [WebNLG]	−0.14	0.14	[−0.41, 0.14]	−0.95	0.340
Train Type [Dat_Aug] × Dataset [CACAPO (en)]	0.10	0.10	[−0.10, 0.30]	0.96	0.335
Train Type [Pseu_Lab] × Dataset [CACAPO (en)]	−0.16	0.10	[−0.37, 0.04]	−1.59	0.111
Train Type [Dat_Aug] × Dataset [CACAPO (nl)]	0.02	0.10	[−0.18, 0.23]	0.23	0.820
Train Type [Pseu_Lab] × Dataset [CACAPO (nl)]	0.10	0.10	[−0.10, 0.30]	0.95	0.342
Train Type [Dat_Aug] × Dataset [WebNLG]	0.06	0.10	[−0.13, 0.26]	0.66	0.508
Train Type [Pseu_Lab] × Dataset [WebNLG]	−0.04	0.10	[−0.23, 0.15]	−0.45	0.656

Appendix D. Qualitative Analysis

CACAPO (NL) – Lowest

Table D.1
Five lowest rated texts per semi-supervised learning type for the Dutch CACAPO dataset, in ascending order.

Train Type	Data	Text
No_Ext	windDirection="zuid-_tot_zuidwestelijke" windAmount="matig"	De zuid- tot zuidwestelijke is matig. *The south- to southwestern is moderate.*
	goalType="rebound" numberOfMatchGoals="beide"	De rebound was een beide keren. *The rebound was a both times.*
	playerName="Klaassen" teamName="Ajacieden" stadiumPlayed="Arnhem"	Klaassen, afgelopen zomer gehuurd van Ajacieden, was in Arnhem de enige speler die de speler maakte. *Klaassen, loaned from Ajacieden past summer, was the only player in Arnhem that made the player.*
	locationArea="In_het_noorden" windDirection="noordoostelijk"	In het noorden is de lucht noordoostelijk. *In the north, the air is northeastern.*
	victimGender1="Hij" victimGender2="Hij" suspectAge="20-jarige" suspectVehicle="automobilist" suspectAddress="Deventer" location="Ceintuurbaan"	Hij, een 20-jarige automobilist uit Deventer, reed over de Schinderbaan in een richting van de richting. *He, a 20-year-old car driver from Deventer, drove on the Schinderbaan in the direction of the direction.*
Dat_Aug	goalType="rebound" numberOfMatchGoals="beide"	De rebound was een beide keren. *The rebound was a both times.*
	companyName1="Ethereum" companyName2="XRP" moneyAmount1="$23,3B" moneyAmount2="$13,5B" stockChangePercentage1="8,66%" stockChangePercentage2="5,01%"	De markt van Ethereum staat inmiddels op $23,3B ofwel 8,66% van de marktwaarde van alle meer, terwijl de allemaal van XRP een volledige van $13,5B bereikte, ofwel 5,1% van de 50 bereikt van de 25. *Ethereum's market is now at $23.3B or 8.66% of the market value of all more, while XRP's all reached a full of $13.5B, or 5.1% of the 50 reached of the 25.*
	timePoint="In_de_avond" locationArea="overal" windChange="af"	In de avond neemt de lucht overal af naar breed tot breed, langs de kust en op het vasteland naar breed tot breed. *In the evening the sky decreases to broad to broad everywhere, along the coast and on the mainland to broad to broad.*
	timePoint="Vanavond" windChange1="af" windChange2="west" locationArea="boven_land"	Vanavond neemt de lucht af en draait boven land naar west. *Tonight the sky will decrease and turn west over land.*
	teamName1="NEC" teamName2="Heracles_Almelo" chanceForName="Navarone_Foor" chanceForType="hard_en_hoog_overschoot" assistType="doorkoppen" assistName="Marcel_Ritzmaier"	NEC kwam eerst in de problemen, maar Navarone Foor was hard en hoog overschoot toen hij een doorkoppen van Marcel Ritzmaier op de lijn belandde. *NEC got into trouble first, but Navarone Foor was hard and high overshooting when he a pass-on heading from Marcel Ritzmaier landed on the line.*
Pseu_Lab	chanceForName1="Schaken" chanceForName2="Timothy_Derijck" matchTime="even_later"	Schaken schoot even later net over, terwijl Timothy Derijck de bal net naast schoot. *Schaken just shot over some time later, while Timothy Derijck shot the ball just wide.*
	goalScore="keerde" teamName="FC_Twente" matchTime="na_vijf_minuten" goalName="Chinedu_Ede"	Het keerde FC Twente al na vijf minuten via Chinedu Ede op het scorebord. *It turned FC Twente after five minutes via Chinedu Ede on the scoreboard.*
	victimGender="man" cause="gestruikeld" incidentType="ongeluk"	Volgens de politie heeft de man gestruikeld en is hij door het ongeluk van de weg geraakt. *Police say the man has tripped and because of the incident went off the road.*
	companyName1="Luchtvaartcombinatie_Air_France-KLM" companyName2="luchtvaartmaatschappij_easyJet" stockChange="min" stockChangePercentage="1,8_procent"	Luchtvaartcombinatie Air France-KLM (min 1,8 procent) en luchtvaartmaatschappij easyJet *Aviation combination Air France-KLM (minus 1.8 percent) and airline easyJet*
	locationArea="overal" weatherType="droog" maximumTemperature="3°C" timePoint="in_de_avond" temperatureCelsius="onder_het_vriespunt"	Het blijft overal droog en de maximumtemperatuur varieert van 3°C in het noordoosten tot +3°C in de avond in het zuiden van het land. *It remains dry everywhere and the maximum temperature varies from 3°C in the northeast to +3°C in the evening in the south of the country.*

CACAPO (NL) – Highest

Table D.2
Five highest rated texts per semi-supervised learning type for the Dutch CACAPO dataset, in descending order.

Train Type	Data	Text
No_Ext	victimGender="vrouw" victimStatus="met_spoed_naar_het_ziekenhuis_overgebracht, _maar_overleed_helaas_in_het_ziekenhuis_aan_haar_ verwondingen"	De vrouw werd met spoed naar het ziekenhuis overgebracht, maar overleed helaas in het ziekenhuis aan haar verwondingen. *The woman was rushed to hospital, but sadly died of her injuries in hospital.*
	timePoint="In_de_avond" weatherChange="geleidelijk" weatherType="droog"	In de avond wordt het geleidelijk droog. *In the evening it gradually becomes dry.*
	incidentType="zware_aanrijding" location="Weimarstraat" dateTime="aan_het_begin_van_de_avond" victimAge="40-jarige" victimVehicle="fietser" victimStatus="om_het_leven_gekomen"	Bij een zware aanrijding op de Weimarstraat is aan het begin van de avond een 40-jarige fietser om het leven gekomen. *A 40-year-old cyclist was killed in a serious collision on the Weimarstraat at the beginning of the evening.*
	teamName="Utrecht" matchTime="na_32_minuten" goalScore="op_gelijke_hoogte"	Utrecht kwam na 32 minuten op gelijke hoogte. *Utrecht equalized after 32 minutes.*
	timePoint="Vannacht" cloudAmount="zwaar_bewolkt" locationArea="langs_de_westkust" weatherIntensity="enkele" weatherType="buien"	Vannacht is het zwaar bewolkt met langs de westkust enkele buien. *Tonight it will be heavily cloudy with a few showers along the west coast.*
Dat_Aug	timePoint="In_de_avond" weatherChange="geleidelijk" weatherType="droog"	In de avond wordt het geleidelijk droog. *In the evening it gradually becomes dry.*
	victimGender="vrouw" victimStatus="met_spoed_naar_het_ziekenhuis_overgebracht, _maar_overleed_helaas_in_het_ziekenhuis_aan_haar_ verwondingen"	De vrouw werd met spoed naar het ziekenhuis overgebracht, maar overleed helaas in het ziekenhuis aan haar verwondingen. *The woman was rushed to hospital, but sadly died of her injuries in hospital.*
	incidentType="zware_aanrijding" location="Weimarstraat" dateTime="aan_het_begin_van_de_avond" victimAge="40-jarige" victimVehicle="fietser" victimStatus="om_het_leven_gekomen"	Bij een zware aanrijding op de Weimarstraat is aan het begin van de avond een 40-jarige fietser om het leven gekomen. *A 40-year-old cyclist was killed in a serious collision on the Weimarstraat at the beginning of the evening.*
	weatherArea="plaatselijk" weatherType="glad"	Het blijft plaatselijk glad. *It remains slippery in places.*
	timePoint="Vannacht" cloudAmount="zwaar_bewolkt" locationArea="langs_de_westkust" weatherIntensity="enkele" weatherType="buien"	Vannacht is het zwaar bewolkt met langs de westkust enkele buien. *Tonight it will be heavily cloudy with a few showers along the west coast.*
Pseu_Lab	minimumTemperature="14°C") windAmount="zwakke_tot_matige" windDirection="zuid_tot_zuidoost"	De minima liggen rond 14°C en er staat een zwakke tot matige wind uit zuid tot zuidoost. *The minima are around 14°C and there is a weak to moderate wind from south to southeast.*
	timePoint="In_de_nacht_naar_donderdag" cloudAmount="veel_bewolking" weatherFrequency="af_en_toe" weatherIntensity="lichte" weatherType="regen"	In de nacht naar donderdag is er veel bewolking, waaruit af en toe lichte regen valt. *In the night to Thursday there will be a lot of clouds, from which light rain occasionally falls.*
	weatherFrequency="perioden" weatherType1="zon" weatherType2="droog" locationArea="op_veel_plaatsen"	Er zijn perioden met zon en het blijft op veel plaatsen droog. *There are periods of sun and it remains dry in many places.*
	weatherType="droog" timePoint="middagtemperatuur" temperatureCelsius="4°C"	Het blijft droog en de middagtemperatuur ligt rond 4°C. *It remains dry and the afternoon temperature is around 4°C.*
	incidentType="zware_aanrijding" location="Weimarstraat" dateTime="aan_het_begin_van_de_avond" victimAge="40-jarige" victimVehicle="fietser" victimStatus="om_het_leven_gekomen"	Bij een zware aanrijding op de Weimarstraat is aan het begin van de avond een 40-jarige fietser om het leven gekomen. *A 40-year-old cyclist was killed in a serious collision on the Weimarstraat at the beginning of the evening.*

CACAPO (EN) – Lowest

Table D.3
Five lowest rated texts per semi-supervised learning type for the English CACAPO dataset, in ascending order.

Train Type	Data	Text
No_Ext	companyName1="KushCo" companyName2="CV_Sciences" timePoint1="September" timePoint2="October"	KushCo a the cannabis - based company, went launched in September following the October of CV Sciences.
	victimNumber="One" victimGender="man" victimStatus="shot_in_the_arm"	One man was shot in the arm, and that gun is still lodged in his thigh.
	locationName="BAHRAIN" stockChange="was_flat" stockPoints="1,584"	BAHRAIN – The stock 's main - largest trading was flat at 1,584 after the on received a positive report from its its.
	victimGender="man" victimStatus="died"	The Medical Department will determine the cause and manner of his died.
	companyName="euro" stockChange="low" timePoint="Tuesday"	The euro hit a low low on Tuesday as investors sought the simplest gain from the crisis 's re.
Dat_Aug	victimNumber1="One" victimNumber2="as_many_as_four" victimStatus="critical_condition"	One of the dead is in critical condition, while as many as four others are in critical condition.
	companyName1="KushCo" companyName2="CV_Sciences" timePoint1="September" timePoint2="October"	KushCo a the cannabis - based company, went launched in September following the October of CV Sciences.
	accidentAddress1="DESOTO_COUNTY,_MS" accidentAddress2="DeSoto_County" victimNumber="Four" victimGender="men" victimStatus="shot"	DESOTO COUNTY, MS (Ben and) – Four men were shot at a house party in DeSoto County on Saturday Sunday.
	locationName="BAHRAIN" stockChange="was_flat" stockPoints="1,584"	BAHRAIN – The stock 's main - largest trading was flat at 1,584 after the on received a positive report from its its.
	takenToHospital="True" hospitalName="Community_Regional_Medical_Center"	The deceased were taken to Community Regional Medical Center with the following injury:
Pseu_Lab	victimNumber1="two" victimNumber2="two"	The other two victims are two men and a woman.
	timePoint1="Monday" timePoint2="Friday_through_Sunday" locationArea="Eugene" temperatureCelsius="60_F."	Monday in Eugene was 60 F. Friday through Sunday.
	pitchResult="homer" pitcherName="Glavine"	The homer chased Glavine.
	timePoint="9pm" temperatureCelsius="-8"	9pm: −8
	companyName="Cimarex_Energy" stockChangePercentage="-4.39%"	Cimarex Energy: (-4.39%)

CACAPO (EN) – Highest

Table D.4
Five highest rated texts per semi-supervised learning type for the English CACAPO dataset, in descending order.

Train Type	Data	Text
No_Ext	exchangeName="S&P_500" stockChange="up" stockPoints="646.78" stockChangePercentage="25.8%"	The S&P 500 is up 646.78 points, or 25.8%.
	exchangeName="Dow" stockChange="up" stockPoints="288.38" stockChangePercentage="1%"	The Dow is up 288.38 points, or 1%.
	exchangeName="S&P_500" stockChange="up" stockPoints="684.29" stockChangePercentage="27.3%"	The S&P 500 is up 684.29 points, or 27.3%.
	stockChange="up" stockChangePercentage="as_much_as_47%" timePoint="this_year"	The company is up as much as 47% this year according to the report.
	exchangeName="S&P_500" stockChange="up" stockPoints="22.34" stockChangePercentage="0.7%"	The S&P 500 is up 22.34 points, or 0.7%.
Dat_Aug	exchangeName="S&P_500" stockChange="up" stockPoints="684.29" stockChangePercentage="27.3%"	The S&P 500 is up 684.29 points, or 27.3%.
	pitcherName="Newcomb" winLossType="won" winLossRecord="seventh"	Newcomb won his seventh consecutive game.
	exchangeName="Russell_2000" stockChange="up" stockPoints="45.16" stockChangePercentage="2.8%"	The Russell 2000 is up 45.16 points, or 2.8%.
	exchangeName="Dow" stockChange="up" stockPoints="288.38" stockChangePercentage="1%"	The Dow is up 288.38 points, or 1%.
	exchangeName="S&P_500" stockChange="up" stockPoints="22.34" stockChangePercentage="0.7%"	The S&P 500 is up 22.34 points, or 0.7%.
Pseu_Lab	exchangeName="S&P_500" stockChange="up" stockPoints="684.29" stockChangePercentage="27.3%"	The S&P 500 is up 684.29 points, or 27.3%.
	timePoint="opened" stockChange="on_a_positive_note"	The market opened on a positive note.
	exchangeName="Dow" stockChange="up" stockPoints="288.38" stockChangePercentage="1%"	The Dow is up 288.38 points, or 1%.
	exchangeName="S&P_500" stockChange="up" stockPoints="646.78" stockChangePercentage="25.8%"	The S&P 500 is up 646.78 points, or 25.8%.
	batterName="Quentin" pitchResult="homered" gameNumber="three"	Quentin has homered in three straight games.

WebNLG – Lowest

Table D.5
Five lowest rated texts per semi-supervised learning type for the WebNLG dataset, in ascending order.

Train Type	Data	Text
No_Ext	"Batchoy" — country — "Philippines" "Philippines" — language — "Philippine Spanish"	Batchoy is eaten in the their, a the of the their spoken is is.
	"Batagor" — country — "Indonesia" "Indonesia" — leaderName — "Jusuf Kalla"	Jusuf Kalla is the his of the the this.
	"11th Mississippi Infantry Monument" — category — "Contributing property" "11th Mississippi Infantry Monument" — location — "Adams County, Pennsylvania"	It is categorized as a contributing site and is located in its.
	"Bionico" — course — "Dessert" "Bionico" — country — "Mexico" "Mexico" — leaderName — "Silvano Aureoles Conejo" "Dessert" — dishVariation — "Cookie"	A type of pie is pie which is a type of pie and is led by Silvano Aureoles Conejo.
	"Amatriciana sauce" — country — "Italy" "Amatriciana sauce" — region — "Lazio"	Amatriciana sauce is a traditional traditional from the Lazio region.
Dat_Aug	"Bhajji" — region — "Karnataka" "India" — demonym — "Indian people" "Bhajji" — country — "India"	Bhajji comes from the this of Karnataka, where the the are called called and are found.
	"11th Mississippi Infantry Monument" — category — "Contributing property" "11th Mississippi Infantry Monument" — location — "Adams County, Pennsylvania"	It is categorized as a contributing site and is located in its.
	"Batagor" — country — "Indonesia" "Indonesia" — leaderName — "Jusuf Kalla"	Jusuf Kalla is the his of the the this.
	"Bionico" — ingredient — "Granola" "Bionico" — course — "Dessert" "Bionico" — region — "Jalisco" "Bionico" — country — "Mexico"	Bionico, which contains contains, is a its found in the this of Jalisco, Mexico.
	"Arrabbiata sauce" — ingredient — "Tomato" "Arrabbiata sauce" — region — "Rome" "Arrabbiata sauce" — country — "Italy"	Arrabbiata sauce, which contains contains, is from the the of Rome, Italy.
Pseu_Lab	"United States" — demonym — "Americans" "United States" — language — "English language" "United States" — ethnicGroup — "Asian Americans"	The American people are a diverse ethnic group, primarily of Asian descent, and are a major part of the English language in the United States.
	"Albany, Oregon" — isPartOf — "Oregon" "Albany, Oregon" — country — "United States" "United States" — capital — "Washington, D.C." "United States" — demonym — "Americans"	Albany is part of Oregon, in the United States, and is the capital of the nation.
	"Spain" — demonym — "Spaniards" "Spain" — leaderName — "Felipe VI of Spain" "Andalusia" — leaderName — "Susana Díaz" "Ajoblanco" — country — "Spain" "Ajoblanco" — region — "Andalusia"	The name is derived from the Spanish word for Spaniards, Felipe VI of Spain, and Susana Diaz of Andalusia, where it is known as a "stamp".
	"Bananaman" — broadcastedBy — "BBC" "BBC" — product — "BBC Radio"	The BBC broadcast the TV series and also produce BBC radio.
	"Belgium" — leaderName — "Charles Michel" "Belgium" — leaderName — "Philippe of Belgium"	Two of the leaders of which are Charles Michel and Philippe of Belgium.

WebNLG – Highest

Table D.6
Five highest rated texts per semi-supervised learning type for the WebNLG dataset, in descending order.

Train Type	Data	Text
No_Ext	"Azerbaijan" — leaderName — "Artur Rasizade" "Azerbaijan" — leaderTitle — "Prime Minister of Azerbaijan"	Artur Rasizade is the Prime Minister of Azerbaijan.
	"20 Fenchurch Street" — floorCount — "34" "20 Fenchurch Street" — floorArea — "62145.3 (square metres)"	20 Fenchurch Street has 34 stories and a total surface of 62145 square metres.
	"Bibbo Bibbowski" — creator — "Jerry Ordway" "Bibbo Bibbowski" — creator — "Marv Wolfman"	Bibbo Bibbowski was created by Jerry Ordway and Marv Wolfman.
	"Baymax" — firstAppearanceInFilm — "Big Hero 6 (film)" "Big Hero 6 (film)" — starring — "Scott Adsit"	Baymax appeared in the film Big Hero 6 which stars Scott Adsit.
	"Auburn, Alabama" — isPartOf — "Lee County, Alabama" "Auburn, Alabama" — isPartOf — "Alabama" "Alabama" — country — "United States"	Auburn is located within Lee County, Alabama.
Dat_Aug	"Bibbo Bibbowski" — creator — "Jerry Ordway" "Bibbo Bibbowski" — creator — "Marv Wolfman"	Bibbo Bibbowski was created by Jerry Ordway and Marv Wolfman.
	"Arrabbiata sauce" — region — "Rome" "Arrabbiata sauce" — country — "Italy" "Arrabbiata sauce" — ingredient — "Olive oil"	Arrabbiata sauce, from the Roma region, Italy, includes includes oil.
	"Azerbaijan" — leaderName — "Artur Rasizade" "Azerbaijan" — leaderTitle — "Prime Minister of Azerbaijan"	Artur Rasizade is the Prime Minister of Azerbaijan.
	"20 Fenchurch Street" — floorCount — "34" "20 Fenchurch Street" — floorArea — "62145.3 (square metres)"	20 Fenchurch Street has 34 stories and a total surface of 62145 square metres.
	"Baymax" — firstAppearanceInFilm — "Big Hero 6 (film)" "Big Hero 6 (film)" — starring — "Scott Adsit"	Baymax appeared in the film Big Hero 6 which stars Scott Adsit.
Pseu_Lab	"Akita Museum of Art" — location — "Akita, Akita" "Akita Museum of Art" — country — "Japan"	The Akita Museum of Art is located in Akita, Akita, Japan.
	"Ampara Hospital" — region — "Ampara District" "Ampara Hospital" — state — "Eastern Province, Sri Lanka" "Ampara Hospital" — country — "Sri Lanka"	Ampara Hospital is located in Ampara District, in the Eastern Province state of Sri Lanka.
	"3Arena" — architect — "Populous (company)" "3Arena" — location — "Dublin"	The company Populous were the architects of the 3Arena in Dublin.
	"250 Delaware Avenue" — floorArea — "30843.8 (square metres)" "250 Delaware Avenue" — floorCount — "12"	250 Delaware Avenue has a floor area of 30843.8 square metres and a floor count of 12.
	"Akita Museum of Art" — location — "Akita, Akita" "Akita, Akita" — isPartOf — "Akita Prefecture" "Akita Museum of Art" — country — "Japan"	The Akita Museum of Art is an art museum in the city of Akita, Akita Prefecture, Japan.

E2E – Lowest

Table D.7
Five lowest rated texts per semi-supervised learning type for the E2E dataset, in ascending order.

Train Type	Data	Text
No_Ext	name="Clowns" eatType="pub" customerRating="3 out of 5" near="All Bar One"	Clowns pub is a local - priced, local - rated, and - friendly - friendly coffee bar located in the city market, near All Bar One.
	name="Giraffe" eatType="restaurant" food="French" area="riverside" familyFriendly="no" near="Raja Indian Cuisine"	Giraffe is a restaurant serving French English in the riverside waterfront. It is not a any - friendly based and is located near the Raja Indian Cuisine Coffee.
	name="The Cricketers" eatType="coffee shop" customerRating="average" familyFriendly="no" near="Crowne Plaza Hotel"	The Cricketers coffee House is a average rated, not company - oriented, sushi cafe located near the Crowne Plaza Hotel.
	name="Zizzi" eatType="pub" customerRating="high" near="Burger King"	Zizzi pub is a local - priced, high rated, and rated rated, located near Burger King.
	name="Giraffe" eatType="pub" food="Fast food" area="riverside" familyFriendly="yes" near="Raja Indian Cuisine"	Giraffe is a pub styled, fast food styled in the riverside area. It is isn fun and located near the Raja Indian Cuisine Coffee.
Dat_Aug	name="Giraffe" eatType="restaurant" food="French" area="riverside" familyFriendly="no" near="Raja Indian Cuisine"	Giraffe is a restaurant serving French English in the riverside waterfront. It is not a any - friendly based and is located near the Raja Indian Cuisine Coffee.
	name="Zizzi" eatType="pub" customerRating="high" near="Burger King"	Zizzi pub is a local - priced, high rated, and rated rated, located near Burger King.
	name="Clowns" eatType="pub" customerRating="3 out of 5" near="All Bar One"	Clowns pub is a local - priced, local - rated, and - friendly - friendly coffee bar located in the city market, near All Bar One.
	name="Wildwood" eatType="restaurant" food="Indian" area="city centre" familyFriendly="no" near="Raja Indian Cuisine"	The Rama is a restaurant that serves Indian meat in the city centre east. It is not Kids oriented and is located near the Raja Indian Cuisine Cafe.
	name="Wildwood" eatType="restaurant" food="Italian" area="riverside" familyFriendly="yes" near="Raja Indian Cuisine"	The Rama is a restaurant serving Italian cuisine in the riverside strip. It is fun - based and located near the Raja Indian Cuisine Café.
Pseu_Lab	name="The Cricketers" eatType="coffee shop" customerRating="high" familyFriendly="yes" near="Cafe Sicilia"	The Cricketers is a restaurant that serves Samoan food. It is located near the Cafe de Bruxelles.
	name="Zizzi" eatType="coffee shop" area="riverside"	Zizzi is a restaurant that serves Samoan food. It is located in the riverside area near the Ying.
	name="The Cricketers" eatType="coffee shop" customerRating="3 out of 5" familyFriendly="yes" near="Crowne Plaza Hotel"	The Cricketers is a restaurant that serves Burmese food. It is located near the Crown Plaza Hotel and is family friendly.
	name="Wildwood" eatType="restaurant" food="Italian" area="city centre" familyFriendly="yes" near="Raja Indian Cuisine"	Wildwood Cafe is a restaurant that serves Saint Helenian food. It is located in the city centre near Raja.
	name="Giraffe" eatType="coffee shop" near="The Rice Boat"	Girafe is a restaurant that serves Burundian food. It is located near The Rice Boat.

E2E – Highest

Table D.8
Five highest rated texts per semi-supervised learning type for the E2E dataset, in descending order.

Train Type	Data	Text
No_Ext	name="Clowns" eatType="coffee shop" near="The Sorrento"	Clowns coffee shop is located near The Sorrento.
	name="Cocum" eatType="coffee shop" customerRating="high" near="Burger King"	Cocum coffee House is a five star cafe located near Burger King.
	name="The Cricketers" eatType="restaurant" customerRating="high" familyFriendly="yes" near="Express by Holiday Inn"	The Cricketers restaurant is a 5 - star family dining located near the Express by Holiday Inn.
	name="The Cricketers" eatType="restaurant" customerRating="3 out of 5" familyFriendly="yes" near="Avalon"	The Cricketers restaurant has a customer grade of 3 out of 5 and is a baby - oriented facility located near Avalon.
	name="The Cricketers" eatType="coffee shop" customerRating="3 out of 5" familyFriendly="yes" near="Crowne Plaza Hotel"	The Cricketers coffee House is rated 3 out of 5 by patrons. It is a baby - oriented cafe located near the Crowne Plaza Hotel.
Dat_Aug	name="Cocum" eatType="coffee shop" customerRating="high" near="Burger King"	Cocum coffee House is a five star cafe located near Burger King.
	name="Giraffe" eatType="pub" near="The Rice Boat"	Giraffe is a pub type style located near The Rice Boat.
	name="Blue Spice" eatType="restaurant" food="English" area="city centre" familyFriendly="yes" near="Rainbow Vegetarian Cafe"	Blue Dragon is a restaurant serving English cooking in the city centre east. It is child - oriented and located near Rainbow Vegetarian Cafe.
	name="Clowns" eatType="coffee shop" near="The Sorrento"	Clowns coffee shop is located near The Sorrento.
	name="The Cricketers" eatType="coffee shop" customerRating="high" familyFriendly="yes" near="Cafe Sicilia"	The Cricketers coffee House is a 5 - star family dining located near Cafe Sicilia.
Pseu_Lab	name="Giraffe" eatType="restaurant" food="French" area="riverside" familyFriendly="no" near="Raja Indian Cuisine"	Girafe is a restaurant that serves French food. It is located in the riverside area near Raja. It is not family friendly.
	name="Blue Spice" eatType="restaurant" food="Chinese" area="riverside" familyFriendly="yes" near="Rainbow Vegetarian Cafe"	Blue Spice is a restaurant that serves Chinese food. It is located in the riverside area near the Rainbow Vegetarian Cafe.
	name="Blue Spice" eatType="restaurant" food="English" area="city centre" familyFriendly="yes" near="Rainbow Vegetarian Cafe"	Blue Spice is a restaurant that serves British food. It is located in the city centre near the Rainbow Vegetarian Cafe.
	name="Giraffe" eatType="pub" food="Fast food" area="city centre" familyFriendly="yes" near="Rainbow Vegetarian Cafe"	Girafe is a family-friendly pub located in the city centre near the Rainbow Vegetarian Cafe.
	name="Blue Spice" eatType="pub" food="Chinese" area="riverside" familyFriendly="yes" near="Rainbow Vegetarian Cafe"	Blue Spice is a family friendly pub located near the Rainbow Vegetarian Cafe.

Appendix E. Ablation Study Graphs

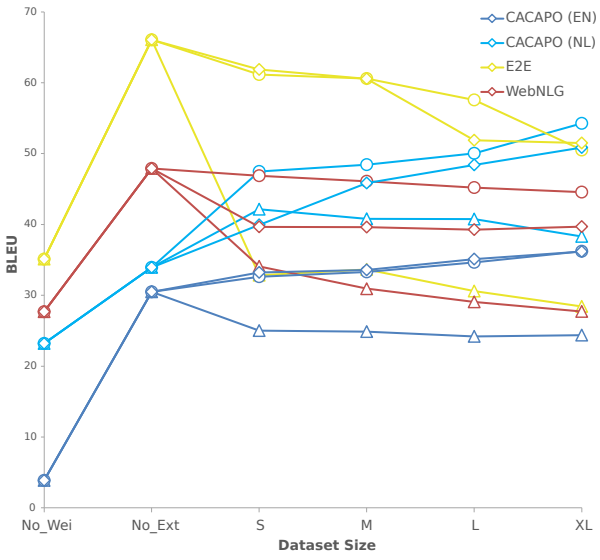

Figure E.1
BLEU scores of the datasets per dataset extension, including no pretrained model and pseudo-labeling and data augmentation combined. Round markers = pseudo-labeling; triangle markers = data augmentation; diamond markers = pseudo-labeling + data augmentation.

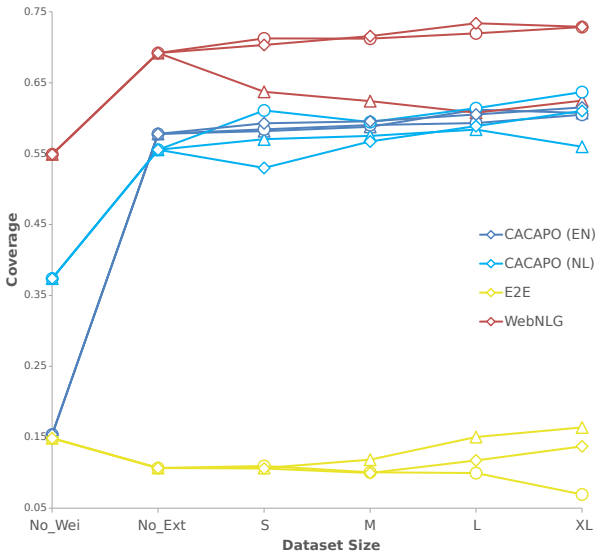

Figure E.2
Coverage scores of the datasets per dataset extension, including no pretrained model and pseudo-labeling and data augmentation combined. Round markers = pseudo-labeling; triangle markers = data augmentation; diamond markers = pseudo-labeling + data augmentation.

Acknowledgments

We received support from RAAK-PRO SIA (2014-01-51PRO) and The Netherlands Organization for Scientific Research (NWO 360-89-050).

References

Agarwal, Oshin, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 125–130.

Ahmadi, Naser, Hansjorg Sand, and Paolo Papotti. 2021. Unsupervised matching of data and text. *arXiv preprint arXiv:2112.08776*. https://doi.org/10.1109/ICDE53745.2022.00084

Alberti, Chris, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173. https://doi.org/10.18653/v1/P19-1620

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

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623. https://doi.org/10.1145/3442188.3445922

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, Curran Associates, Inc.

Burke, Robin D., Kristian J. Hammond, and B. C. Yound. 1997. The FindMe approach to assisted browsing. *IEEE Expert*, 12(4):32–40. https://doi.org/10.1109/64.608186

Castro Ferreira, Thiago, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76.

Castro Ferreira, Thiago, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176. https://doi.org/10.18653/v1/W18-6521

Castro Ferreira, Thiago, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562. https://doi.org/10.18653/v1/D19-1052

Castro Ferreira, Thiago, Helena Vaz, Brian Davis, and Adriana Pagano. 2021. Enriching the E2E dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 177–183.

Chang, Ernie, Vera Demberg, and Alex Marin. 2021. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 818–829. https://doi.org/10.18653/v1/2021.eacl-main.69

Chang, Ernie, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui Su. 2021. Neural data-to-text generation with LM-based text augmentation. In *Proceedings of the 16th Conference of the European Chapter of the*

Association for Computational Linguistics: Main Volume, pages 758–768. https://doi.org/10.18653/v1/2021.eacl-main.64

Chen, Wenhu, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648. https://doi.org/10.18653/v1/2020.emnlp-main.697

Clerwall, Christer. 2014. Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, 8(5):519–531. https://doi.org/10.1080/17512786.2014.883116

Davies, Mark and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051. https://doi.org/10.2307/2529886

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145. https://doi.org/10.3115/1289189.1289273

Emmery, Chris, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2388–2402. https://doi.org/10.18653/v1/2021.eacl-main.203

Emmery, Chris, Ákos Kádár, Grzegorz Chrupala, and Walter Daelemans. 2022. Cyberbullying classifiers are sensitive to model-agnostic perturbations. *arXiv preprint arXiv:2201.06384*.

Faille, Juliette, Albert Gatt, and Claire Gardent. 2020. The natural language pipeline, neural text generation and explainability. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 16–21.

Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188. https://doi.org/10.18653/v1/P17-1017

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133. https://doi.org/10.18653/v1/W17-3518

Gatt, Albert and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. https://doi.org/10.1613/jair.5477

Gkatzia, Dimitra. 2016. Content selection in data-to-text systems: A survey. *arXiv preprint arXiv:1610.08375*.

Greco, Claudio, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605. https://doi.org/10.18653/v1/P19-1350

He, Junxian, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Heidari, Peyman, Arash Einolghozati, Shashank Jain, Soumya Batra, Lee Callender, Ankit Arun, Shawn Mei, Sonal Gupta, Pinar Donmez, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Getting to production with few-shot natural language generation models. In

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 66–76.
Hendriks, Barry. 2019. Retrieving, Cleaning and Analysing Dutch news articles about traffic accidents. Master's thesis, University of Amsterdam, The Netherlands.
Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*, pages 1–16.
Honnibal, Matthew and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://doi.org/10.5281/zenodo.3358113
Hoorn, Johan F. and Teunis D. van Wijngaarden. 2010. Web intelligence for the assessment of information quality: Credibility, correctness, and readability, Usmani, Zeeshan Ul Hassan, editor, *Web Intelligence and Intelligent Agents*. InTech, pages 205–324.
Jolly, Shailza, Zi Xuan Zhang, Andreas Dengel, and Lili Mou. 2022. Search and learn: Improving semantic coverage for data-to-text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10858–10866. https://doi.org/10.1609/aaai.v36i10.21332
Kale, Mihir and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.
Kedzie, Chris and Kathleen McKeown. 2019. A good sample is hard to find: Noise injection sampling and self-training for neural language generation models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 584–593. https://doi.org/10.18653/v1/W19-8672
Kulhánek, Jonáš, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210. https://doi.org/10.18653/v1/2021.nlp4convai-1.19
Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. https://doi.org/10.2307/2529310
Lin, Chin Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
Mager, Manuel, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852. https://doi.org/10.18653/v1/2020.acl-main.167
McCarthy, Diana and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53. https://doi.org/10.3115/1621474.1621483
Mehta, Sanket Vaibhav, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur Parikh, and Emma Strubell. 2022. Improving compositional generalization with self-training for data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4205–4219. https://doi.org/10.18653/v1/2022.acl-long.289
Montella, Sebastien, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Rojas-Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced RDF verbalization with transformers. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 89–99.
Nguyen, Minh Tien, Viet-Anh Phan, Le Thai Linh, Nguyen Hong Son, Le Tien Dung, Miku Hirano, and Hajime Hotta. 2019. Transfer learning for information extraction with limited data. In *International Conference of the Pacific Association for Computational Linguistics*, pages 469–482. https://doi.org/10.1007/978-981-15-6168-9_38
Novikova, Jekaterina, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206. https://doi.org/10.18653/v1/W17-5525
Oraby, Shereen, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and

method for joint control of semantics and style in neural NLG. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5938–5951. https://doi.org/10.18653/v1/P19-1596

Papineni, Kishore, 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. https://doi.org/10.3115/1073083.1073135

Parikh, Ankur, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. https://doi.org/10.18653/v1/2020.emnlp-main.89

Puduppully, Ratish, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035. https://doi.org/10.18653/v1/P19-1195

Qader, Raheel, François Portet, and Cyril Labbé. 2019. Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 552–562. https://doi.org/10.18653/v1/W19-8669

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Riabi, Arij, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030. https://doi.org/10.18653/v1/2021.emnlp-main.562

Ribeiro, Leonardo F. R., Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227. https://doi.org/10.18653/v1/2021.nlp4convai-1.20

Rizos, Georgios, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 991–1000. https://doi.org/10.1145/3357384.3358040

Roberti, Marco, Giovanni Bonetta, Rossella Cancelliere, and Patrick Gallinari. 2020. Copy mechanism and tailored training for character-based data-to-text generation. In *Machine Learning and Knowledge Discovery in Databases*, pages 648–664. https://doi.org/10.1007/978-3-030-46147-8_39

Ross, John Robert. 1979. Where's English? In Daniel Kempler, Charles J. Fillmore, and William S.-Y. Wang, editors, *Individual Differences in Language Ability and Language Behavior*, pages 127–163. https://doi.org/10.1016/B978-0-12-255950-1.50014-7

Schmitt, Martin, Sahand Sharifzadeh, Volker Tresp, and Hinrich Schütze. 2020. An unsupervised joint system for text generation from knowledge graphs and semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7117–7130. https://doi.org/10.18653/v1/2020.emnlp-main.577

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Shimanaka, Hiroki, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758. https://doi.org/10.18653/v1/W18-6456

Shimorina, Anastasia, Elena Khasanova, and Claire Gardent. 2019. Creating a corpus for Russian data-to-text generation using neural machine translation and post-editing. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 44–49. https://doi.org/10.18653/v1/W19-3706

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan

Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. https://doi.org/10.18653/v1/P19-1355

Su, Shang Yu, Chao-Wei Huang, and Yun-Nung Chen. 2020. Towards unsupervised language understanding and generation by joint dual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 671–680. https://doi.org/10.18653/v1/2020.acl-main.63

Sun, Zijun, Chun Fan, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Neural semi-supervised learning for text classification under large-scale pretraining. *arXiv preprint arXiv:2011.08626*.

Sundar, S. Shyam. 1999. Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly*, 76(2):373–386. https://doi.org/10.1177/107769909907600213

Surya, Sai, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068. https://doi.org/10.18653/v1/P19-1198

Tandon, Shubhangi, T. S. Sharath, Shereen Oraby, Lena Reed, Stephanie Lukin, and Marilyn Walker. 2018. TNT-NLG, System 2: Data repetition and meaning representation manipulation to improve neural generation. *E2E NLG Challenge System Descriptions*, pages 1–8.

Tseng, Bo Hsiang, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1795–1807. https://doi.org/10.18653/v1/2020.acl-main.163

van der Lee, Chris, Chris Emmery, Sander Wubben, and Emiel Krahmer. 2020. The CACAPO dataset: A multilingual, multi-domain dataset for neural pipeline and end-to-end data-to-text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 68–79.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151. https://doi.org/10.1016/j.csl.2020.101151

van Miltenburg, Emiel, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741.

van Miltenburg, Emiel, Merel van de Kerkhof, Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. 2019. On task effects in NLG corpus elicitation: A replication study using mixed effects modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 403–408. https://doi.org/10.18653/v1/W19-8649

van Miltenburg, Emiel, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623. https://doi.org/10.18653/v1/2021.naacl-main.51

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Wang, Hongmin. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322. https://doi.org/10.18653/v1/W19-8639

Wiseman, Sam, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263. https://doi.org/10.18653/v1/D17-1239

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.

Zhang, Ruixue, Wei Yang, Luyun Lin, Zhengkai Tu, Yuqing Xie, Zihang Fu, Yuhao Xie, Luchen Tan, Kun Xiong, and Jimmy Lin. 2020a. Rapid adaptation of BERT for information extraction on domain-specific business documents. *arXiv preprint arXiv:2002.01861*.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations*, pages 1–43.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.

Zhao, Wei, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578. https://doi.org/10.18653/v1/D19-1053

Zhao, Yanbin, Lu Chen, Zhi Chen, and Kai Yu. 2020. Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9668–9675. https://doi.org/10.1609/aaai.v34i05.6515

Zhou, Wangchunshu, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373. https://doi.org/10.18653/v1/P19-1328