UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

# Multilingual Graph-to-Text Generation and Evaluation

# THÈSE

présentée et soutenue publiquement le 29 Septembre 2025

pour l'obtention du

**Doctorat de l'Université de Lorraine**

**(Mention Informatique)**

par

William Eduardo Soto Martinez

**Composition du jury**

| | | |
|---|---|---|
| *Président :* | | |
| *Rapporteurs :* | François Portet | Professeur des Universités<br>Université Grenoble Alpes, France |
| | François Yvon | Directeur de Recherche<br>CNRS, ISIR, Sorbonne Université, France |
| *Examinateurs :* | Maxime Amblard | Professeur des Universités<br>Université de Lorraine, France |
| | Oana Balalau | Chargée de Recherche (ISPF)<br>Centre Inria de Saclay, Inria, France |
| | Ondrej Ducek | Maître de Conférence<br>Charles University, Tchéquie |
| | Simon Mille | Chargé de Recherche (Marie Curie)<br>Dublin City University, Irlande |
| *Directrice de thèse :* | Claire Gardent | Directrice de Recherche<br>CNRS, Loria, Université de Lorraine, France |
| *Co-directeur de thèse :* | Yannick Parmentier | Maître de Conférences<br>Université de Lorraine, France |

Mis en page avec la classe thesul.

# Acknowledgements

Successfully conducting research and writing a thesis are no easy feats, and this accomplishment would not have been possible without the constant support and contributions of many people. Science is, after all, a communal endeavor, and I am deeply grateful to everyone who made this work possible in one way or another.

First and foremost, I want to thank my supervisor, Claire Gardent, and my co-supervisor, Yannick Parmentier. Their guidance, expertise, and support have been invaluable throughout my journey. I am incredibly grateful for their trust in me and for the opportunities they provided, which have shaped me into the researcher I am today.

I would also like to thank the members of my defense committee for reading and reviewing my work, for their thoughtful feedback, and for their presence and comments during my defense.

Another heartfelt thank-you goes to my colleagues in the research team. I have been fortunate to work alongside many talented individuals, from those who graduated when I had just started to those who joined as I am finishing. Together, we sparked scientific curiosity, exchanged ideas, and built a support system that extended beyond our academic pursuits. In particular, I want to thank Kelvin Han for his constant support. Whether through insightful discussions, practical assistance, or everyday advice, his help has been truly invaluable.

Lastly, I want to thank my friends and family, whose encouragement and understanding helped me persevere, especially during the most challenging moments. In particular, I would like to thank my parents for their unwavering encouragement and my life partner, Caroline, for her endless love, patience, and unwavering support. I cannot express enough how much their belief in me gave me the resilience to keep moving forward.

*To my parents and Caroline:*
*Without your love and support, I would not be here.*

# Abstract

The efficient communication of structured knowledge is a longstanding challenge in Natural Language Processing (NLP), particularly for Natural Language Generation (NLG). Structured data, such as Resource Description Framework (RDF) graphs and Abstract Meaning Representation (AMR) graphs, enables machines to represent knowledge with clarity and consistency. However, natural language remains the most effective medium for human understanding. This thesis advances the Graph-to-Text (G2T) generation task by improving the fluency and semantic faithfulness of text generated from structured graphs in both high- and low-resource languages.

The primary obstacle addressed in the following research is the scarcity of parallel graph-text data, especially for low-resource languages, which hampers the development and evaluation of multilingual G2T systems. To mitigate this, the thesis proposes two strategies that exploit phylogenetic (language-family) information to guide model training. The first strategy introduces monolingual denoising pre-training with phylogeny-informed soft prompts, followed by full fine-tuning, to improve RDF-to-Text generation in low-resource Celtic languages. The second strategy presents a multilingual framework for AMR-to-Text generation that combines synthetic training data with a hierarchical curriculum of Quantized Low-Rank Adapters (QLoRA), also driven by phylogenetic information. Both methods deliver consistent gains in generation quality, particularly in languages with limited labeled data, by maximizing cross-lingual transfer while controlling training noise.

Beyond generation, the thesis examines current evaluation methodologies. Acknowledging the limitations of reference-based metrics, especially in under-resourced languages. In that regard, this thesis proposes a reference-less metric for assessing RDF-to-Text generation. Leveraging Natural Language Inference (NLI), the metric directly measures semantic faithfulness between graphs and generated texts, providing semantic precision, recall, and F1 figures across diverse languages.

Collectively, these contributions advance the inclusivity and reliability of multilingual G2T generation and evaluation. By addressing data scarcity through phylogenetic transfer and designing principled evaluation frameworks, this research contributes to the democratization of language technology and promotes equitable access to structured knowledge.



**Keywords:** Natural Language Generation (NLG), Graph-to-Text (G2T), Multilingual, Low-Resource (LR) Languages, Phylogenetic Transfer, Semantic Faithfulness, Reference-less Evaluation, Natural Language Inference (NLI).

# Résumé

Communiquer efficacement des connaissances structurées demeure un défi majeur du traitement automatique des langues (TAL), en particulier dans le contexte de la génération automatique de texte (GAT). Bien que les données structurées telles que les graphes de type RDF (*Resource Description Framework*) utilisés dans le Web sémantique, et les graphes de type AMR (*Abstract Meaning Representation*) utilisés pour la représentation du sens d'énoncés permettent aux machines de représenter les connaissances avec clarté et cohérence, le langage naturel reste le moyen le plus adapté à la communication humaine. Cette thèse vise à faire progresser la génération de texte à partir de graphes (*Graph-to-Text, G2T*) en améliorant la fluidité et la fidélité sémantique des textes générés dans les langues bien ou moins bien dotées.

Le principal obstacle traité dans ce travail est celui de la rareté des données parallèles graphe-textes, surtout pour les langues dites peu dotées, ce qui freine le développement et l'évaluation de systèmes *G2T* multilingues. Pour y remédier, cette thèse propose deux stratégies exploitant l'information phylogénétique (famille de langue) afin de guider l'apprentissage. La première stratégie introduit un pré-apprentissage monolingue de débruitage avec des prompts "souples" (soft prompts) incluant des informations phylogénétiques, suivi d'un ajustement (fine-tuning) complet, pour améliorer la génération de texte à partir de graphes *RDF* dans des langues celtiques sous-représentées. La deuxième stratégie présente un cadre multilingue de génération de texte à partir de graphes *AMR* reposant sur des données d'entraînement synthétiques, et d'adaptation de faible rang quantifiée (*Quantized Low-Rank Adapters, QLoRA*) également guidée par de l'information phylogénétique. Les deux approches montrent des gains constants de qualité, en particulier dans les langues pour lesquelles les données annotées sont limitées, ceci grâce à un transfert interlinguistique optimisé et à une maîtrise du bruit d'entraînement.

Au-delà de la génération, la thèse s'intéresse également aux méthodes d'évaluation actuelles, et cherche à dépasser les limites des métriques basées sur la référence, en particulier dans le cas des langues sous-dotées. À cet égard, cette thèse propose une métrique multilingue et sans référence pour l'évaluation de textes générés à partir de graphes RDF. En s'appuyant sur des techniques d'inférence en langage naturel (*Natural Language Inference, NLI*), celle-ci mesure directement la fidélité sémantique entre le graphe d'entrée et le texte généré sous forme de précision, rappel et F1 sémantiques, pour diverses langues.

Ces contributions améliorent collectivement l'inclusivité et la fiabilité de la génération et de l'évaluation *G2T* multilingues. En abordant la question de la rareté des données via le transfert phylogénétique et en proposant des cadres d'évaluation fondés sur des principes solides, ce travail soutient la démocratisation des technologies de traitement de la langue et un accès équitable à la connaissance structurée.



**Mots-clés:** Génération automatique de texte (GAT), Génération de texte à partir de graphes, Multilingue, Langues peu doteé, Transfert phylogénétique, Fidélité sémantique, Évaluation sans référence, Inférence en langage naturel (*NLI*).

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AGENDA | Abstract GENeration DAtaset |
| AGTPS | Asymmetric Generalized Traveling Salesman Problem |
| AMR | Abstract Meaning Representation |
| B20 | BLEURT-20 |
| BART | Bidirectional and Auto-Regressive Transformer |
| BCE | Binary Crossentropy Loss |
| BERT | Bidirectional Encoder Representations from Transformers |
| BiRNN | Bidirectional Recurrent Neural Network |
| BLEU | BiLingual Evaluation Understudy |
| BLEURT | BLEU beRT |
| BMR | BabelNet Meaning Representation |
| BP | Brevity Penalty |
| CLM | Causal Language Modeling |
| D2T | Data-to-Text |
| DART | DAta-Record-to-Text |
| DCGCN | Densely Connected Graph Convolutional Network |
| DQE | Data-QuestEval |
| DLH | Distant Language Hierarchy |
| DLHQL | Distant Language Hierarchical QLoRA |
| E2E | End-to-End |
| FFT | Full Fine-Tuning |
| FNN | Feed-forward Neural Network |
| FS | FactSpotter |
| G2T | Graph-to-Text |
| GCN | Graph Convolutional Networks |
| GGNN | Gated Graph Neural Networks |
| GPT | Generative Pre-trained Transformer |
| GPU | Graphics Processing Unit |
| HQL | Hierarchical QLoRA |
| HMM | Hidden Markov Models |
| HR | High-Resource |
| IE | Information Extraction |
| K | Key |
| KELM | Knowledge Enhanced Language Model |
| KG | Knowledge Graph |
| LABSE | Language-agnostic BERT Sentence Embeddings |
| LID | Language IDentification |

| | |
|---|---|
| LM | Language Modeling |
| LLM | Large Language Model |
| LoRA | Low-Rank Adapters |
| LR | Low-Resource |
| LSTM | Long Short-Term Memory |
| mBART | Multilingual BART |
| MLM | Masked Language Modeling |
| MonoQL | Monolingual QLoRA |
| MR | Medium-Resource |
| MRS | Minimal Recursion Semantics |
| MT | Machine Translation |
| mT5 | Multilingual T5 |
| MTL | Multi-Task Learning |
| MultiQL | Multilingual QLoRA |
| NER | Named Entitiy Recognition |
| NLG | Natural Language Generation |
| NLI | Natural Language Inference |
| NLLB | No Language Left Behind |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| NMT | Neural Machine Translation |
| NN | Neural Network |
| OWL | Web Ontology Language |
| PI | Phylogeny Inspired |
| PLM | Pre-trained Language Models |
| POS | Part of Speech |
| PTH | Phylogenetic Tree Hierarchy |
| PTHQL | Phylogenetic Tree Hierarchical QLoRA |
| Q | Query |
| QA | Question Answering |
| QL | Quantized Low-Rank Adapters |
| QLoRA | Quantized Low-Rank Adapters |
| QCET | Quality Criteria for Evaluation of Text |
| RDF | Resource Description Framework |
| RMSE | Root Mean Square Error |
| RNN | Recurrent Neural Network |
| SBERT | Sentence-BERT |
| Seq2Seq | Sequence-to-Sequence |
| SLM | Small Language Models |
| SLOR | Syntactic Log-Odds Ratio |
| SMT | Statistical Machine Translation |
| SPL | Sentence Plan Language |
| STILT | Supplementary Training on Intermediate Labeled data Tasks |
| STS | Semantic Textual Similarity |
| ST | Source-Target |
| T2T | Text-to-Text |
| T5 | Text-to-Text Transfer Transformer |
| TER | Translation Edit Rate |

| | |
|---|---|
| TekGen | Text from KG Generator |
| TST | Task-Source-Target |
| TT | Task-Target |
| UMR | Uniform Meaning Representation |
| V | Value |
| W3C | World Wide Web Consortium |
| WPSLOR | WordPiece Syntactic Log-Odds Ratio |

# Génération et Évaluation de Textes Multilingues à partir de Graphes

À travers l'Histoire, les sociétés se sont appuyées sur des données structurées pour capturer, stocker et diffuser le savoir, allant des catalogues antiques jusqu'aux vastes graphes de connaissances actuels. À l'ère numérique, les représentations fondées sur les graphes telles que le Cadre de Description de Ressources (*Resource Description Framework, RDF*) et la Représentation Sémantique Abstraite (*Abstract Meaning Representation, AMR*) sont devenues des outils essentiels pour organiser l'information de façon précise, non ambiguë et exploitable par les machines. Pourtant, aussi utiles soient-elles pour les systèmes informatiques, ces structures ne sont pas naturellement adaptées à l'usage humain. À l'inverse, le langage naturel demeure notre mode de communication le plus expressif et accessible, bien qu'ambigu et difficile à interpréter automatiquement.

La génération de texte à partir de graphes (*Graph-to-Text, G2T*) se situe à cette intersection entre connaissances structurées et langage naturel. Elle vise à traduire la connaissance encodée sous forme de graphes en langage naturel fluide et fidèle sur le plan sémantique (le sens de l'information encodée dans le graphe doit être restitué fidèlement). Les enjeux sont majeurs : alors que le volume de données structurées explose, alimenté par des initiatives comme Wikidata ou Google Knowledge Graph, le besoin de rendre ces informations accessibles dans une grande majorité de langues s'intensifie.

Malgré les avancées, le domaine *G2T* reste largement centré sur l'anglais et d'autres langues qualifiées de bien dotées. Ce n'est pas faute de nécessité, des centaines de millions de locuteurs de langues peu dotées restent exclus des avancées du domaine, mais en raison de la rareté de paires graphe-texte de qualité, indispensables pour l'apprentissage et l'évaluation des systèmes de G2T. Le résultat est une fracture numérique persistante qui prive les langues peu dotées des bénéfices de la génération automatiques de connaissances, accentuant les inégalités globales.

Cette thèse s'inscrit dans la volonté de démocratiser l'accès au savoir structuré. Elle relève deux défis imbriqués : d'une part, la rareté des données et des méthodes pour le *G2T* dans les langues peu dotées, d'autre part l'inadéquation des métriques d'évaluation actuelles, qui dépendent largement de textes de référence souvent inexistants pour une majorité de langues. Répondre à ces défis est à la fois une question technique, sociale et linguistique, afin que les progrès de l'accessibilité aux connaissances par l'IA bénéficient à toutes les communautés linguistiques.

## Objectifs et questions de recherche

Pour dépasser ces limitations, ce travail exploite la phylogénie linguistique, l'étude des relations entre langues, pour orienter l'apprentissage multilingue des modèles *G2T* et l inférence en langage naturel (*Natural Language Inference, NLI*) pour l'évaluation. L'objectif est d'améliorer la génération de texte à partir de données structurées et l'évaluation de celle-ci en facilitant le transfert entre langues apparentées et en réduisant la dépendance aux références de qualité. L'ambition centrale est de promouvoir un *G2T* inclusif, qui puisse passer à une certaine échelle tout en restant fidèle sémantique, cela sur un spectre linguistique élargi.

Ce travail s'articule autour de trois questions de recherche principales :

### QR1. La génération de texte à partir de graphes *RDF* dans les langues peu dotées peut-elle être améliorée par l'adaptation d'un modèle multilingue avec des prompts souples (soft prompts) enrichis d'information phylogénétique ?

Cette question s'inscrit dans le contexte de la pénurie de données pour les langues peu dotées. Elle examine si des soft prompts structurés, encodant la tâche et la famille linguistique, peuvent favoriser un transfert efficace dans trois langues celtiques : le breton, l'irlandais et le gallois.

### QR2. La connaissance phylogénétique peut-elle guider l'entraînement multilingue en génération de texte à partir de graphes *AMR*, tant pour les langues bien que peu dotées ?

Ici, la thèse étend son analyse aux entrées *AMR* et à douze langues indo-européennes. Elle explore si l'apprentissage par curriculum structuré selon les relations entre langues, combiné à une adaptation modulaire, améliore la qualité G2T dans des contextes à ressources variables.

### QR3. L'Inférence en Langage Naturel (*NLI*) peut-elle être utilisée pour définir une métrique d'évaluation sans référence, multilingue, mesurant la fidélité sémantique en génération de texte à partir de graphes *RDF*?

Cette question vise une évaluation qui puisse passer à l'échelle et qui soit indépendante de la langue, en utilisant la *NLI* pour mesurer le recouvrement sémantique entre graphes et textes générés, sans recourir à des références coûteuses à produire.

Ces axes visent ensemble à élargir l'accès au savoir structuré via des techniques *G2T* plus inclusives, transférables et robustes.

## Plan de la thèse

La thèse est construite autour de plusieurs axes de recherche interconnectés, chacun répondant à un enjeu clé de la génération et de l'évaluation *G2T* multilingues. Elle progresse du cadre théorique vers les solutions concrètes.

Le Chapitre 1 présente le sujet, les questions de recherche et la structure de la thèse.

Le Chapitre 2 propose un état de l'art de la génération automatique de texte (GAT), en se concentrant sur *G2T*, les représentations graphiques, les défis multilingues et les pratiques d'évaluation, préparant le terrain pour les contributions de la thèse.

Le Chapitre 3 traite la QR1 en introduisant une nouvelle méthode pour la génération de texte à partir de graphes *RDF* dans les langues celtiques peu dotées. La méthode *PI-TST* (*Phylogeny-Inspired Task-Source-Target Soft Prompts*) combine un pré-entraînement multilingue avec des prompts structurés par relations linguistiques, permettant un transfert interlinguistique avec peu de données.

Le Chapitre 4 répond à la QR2 en appliquant des curricula phylogénétiques et une adaptation modulaire à la génération de texte à partir de graphes *AMR*. Le cadre hiérarchique d'adaptateurs quantifiés de bas rang (*Hierarchical Quantized Low-Rank Adapters, HQL*), efficace en paramètres, adapte progressivement les modèles multilingues aux contextes spécifiques à chaque langue, évalué sur douze langues indo-européennes.

Le Chapitre 5 aborde la QR3 via une métrique d'évaluation sans référence fondée sur la *NLI*. Cette métrique calcule la précision, le rappel et le F1 score sémantiques directement à partir des graphes *RDF* et des textes générés, permettant une évaluation extensible, interprétable et indépendante de la langue.

Le Chapitre 6 synthétise les résultats, discute les limites et esquisse des perspectives de recherches complémentaires. Il réaffirme le fil conducteur de la thèse : démocratiser l'accès au savoir structuré grâce à des systèmes *G2T* multilingues fidèles, efficaces et inclusifs.

## Méthodes et contributions

Cette section contient un bref résumé de la motivation, de la méthodologie et des résultats concernant chaque question de recherche.

### Avancées de la génération de texte à partir de graphes *RDF* pour les langues celtiques

Le premier axe technique vise la génération de texte à partir de graphes *RDF* dans les langues celtiques à faibles ressources. Alors que les jeux de données de grande taille et les modèles puissants ont fait progresser le *G2T* pour les langues bien dotées, la rareté des données reste un obstacle majeur pour les autres langues. Cette thèse pose l'hypothèse que la proximité linguistique, les langues partageant des caractéristiques de famille ou de structure, peut soutenir le transfert interlinguistique. Elle analyse si des soft prompts structurés, informés par la phylogénie, peuvent renforcer les modèles multilingues en contexte de sous-représentation.

L'approche proposée introduit les soft prompts *PI-TST* : des prompts modulaires et structurés qui codent la tâche, la famille, le genre et la langue pour la source et la cible. Ceux-ci sont associés à une base multilingue pré-entraînée ($mT5_{large}$). L'apprentissage se déroule en trois étapes : adaptation du modèle de base par masquage sur corpus monolingue et données *RDF* ; pré-apprentissage non supervisé des prompts via des tâches comme la modélisation du langage ou le *deshuffling* ; et enfin réglage fin sur des petits ensembles *RDF*-Texte.

Les résultats expérimentaux montrent que *PI-TST* surpasse l'adaptation complète du modèle et des techniques de base telles que *Control Prefixes*, avec des gains conséquents sur les métriques automatiques (*BLEU*, *Google BLEU*, cosinus *LaBSE*) et par rapport aux évaluations humaines (lisibilité, grammaticalité, ordre des mots). Les gains les plus notables sont observés pour le breton, langue absente du pré-entraînement, validant l'hypothèse du transfert phylogénétique.

Les études d'ablation démontrent que prompts source et cible sont nécessaires, et que la méthode reste efficiente avec aussi peu que 1 000 exemples par langue.

### Curriculum hiérarchique pour la génération de texte multilingue à partir de graphes *AMR*

Le second axe aborde la génération de texte à partir de graphes *AMR* sur un éventail plus large de langues indo-européennes, bien ou moins bien dotées. Le défi est double : manque de données annotées et risque que l'entraînement multilingue introduise du bruit.

Pour gérer ce compromis entre manque de données et bruitage, la thèse propose *HQL*, une stratégie d'apprentissage par curriculum qui affine itérativement un modèle multilingue en modèles monolingues via des adaptations efficaces en paramètres. Sur une base $mT5_{large}$ quantifiée à 4 bits, les daptateurs de bas rang (*Low-Rank Adapters, LoRA*) permettent un entraînement modulaire avec de faibles coûts en mémoire et données. La hiérarchie suit : un modèle global (L0) affiné en groupes de 6 langues (L1), puis bilingues (L2), puis monolingues (L3), chaque adaptateur *LoRA* étant réutilisé et étendu à chaque étape, réduisant les coûts d'entraînement.

Deux curricula sont explorés : l'un maximisant la diversité linguistique (*Distant Language Hierarchy, DLH*), l'autre regroupant les langues par similarité (*Phylogenetic Tree Hierarchy, PTH*). Les résultats montrent que les regroupements phylogénétiques sont souvent les plus efficaces, tout en préservant l'effet de régularisation du mélange. Les deux curricula dépassent généralement les modèles monolingues et multilingues de référence, ainsi qu'un pipeline Générer-et-Traduire, surtout pour des langues peu dotées comme l'asturien ou le créole haïtien, mais aussi dans certains contextes de langues à fortes ressources.

### Évaluation sans référence : fidélité sémantique multilingue via *NLI*

Le troisième axe concerne la problématique de l'évaluation. La plupart des métriques *G2T* (*BLEU*, *ChrF++*) dépendent de références de qualité, rares hors de quelques langues. Même les métriques sans référence récentes (*Data-QuestEval*, *Factspotter*) restent très centrées sur l'anglais et d'utilité diagnostique limitée.

La thèse propose une métrique sans référence basée sur la *NLI*. À partir d'un modèle *mDeBERTa-v3* multilingue adapté à la *NLI* puis à la régression, la méthode estime la précision sémantique (texte sous-entendu par le graphe), le rappel (graphe sous-entendu par le texte), et leur F1, entre le graphe *RDF* d'entrée et le texte généré. Les données d'entraînement comprennent 1,77 million de paires synthétiques graphe-texte en six langues, créées par manipulation des exemples réels et traduction automatique, avec filtrage pour similarité sémantique et identité linguistique. Les adaptations complètes et *LoRA* sont évaluées.

Les résultats montrent une forte corrélation entre la métrique proposée, les jugements humains et les métriques de référence, bien qu'aucune référence dorée (gold-standard) ne soit requise. La version *LoRA* monolingue surpasse les métriques sans référence dans toutes les langues, atteignant des corrélations de Spearman jusqu'à 0,70 en anglais et 0,67 en russe. La décomposition en scores de précision et de rappel permet un diagnostic précis de la sur- ou sous-génération, précieux pour les langues peu dotées et les développements futurs.

## Limites et perspectives

Malgré les avancées, certaines limites persistent :

- La recherche s'appuie sur des données synthétiques pour les langues peu dotées, qui ne reflètent pas la diversité linguistique ou structurelle, et peuvent introduire des biais.
- La couverture linguistique, bien qu'élargie, reste limitée aux familles indo-européennes. L'efficacité pour des langues typologiquement différentes reste à explorer.
- La métrique sans référence ne permet pas encore une localisation fine des erreurs, ce qui limite son utilité pour le diagnostic détaillé.

Dans ce contexte, les perspectives de travail incluent l'amélioration de la qualité et de la diversité des ressources de données, l'extension de l'expérimentation à d'autres familles et phénomènes linguistiques, et le développement de méthodes d'évaluation plus interprétables et granulaires.

## Conclusion

Cette thèse présente des méthodes robustes pour la génération et l'évaluation multilingues *G2T*, notamment pour les langues à faibles ressources. En exploitant la structure linguistique et des techniques d'apprentissage efficaces, elle fait progresser l'état de l'art pour la génération à partir de graphes *RDF* et *AMR*. Trois contributions majeures se distinguent : des soft prompts phylogénétiques pour la génération de texte en langues celtiques à partir de graphes *RDF*, un curriculum hiérarchique pour la génération à partir de graphes *RDF* en langues indo-européennes, et une métrique sans référence basée sur la *NLI* pour évaluer la fidélité sémantique.

Ces méthodes illustrent comment l'expertise linguistique et l'adaptation ciblée peuvent compenser le manque de données et le bruit multilingue, rendant le *G2T* plus accessible et équitable. Les implications dépassent la recherche : alors que les applications multilingues deviennent centrales dans la société numérique, les techniques développées ici offrent une base pour des technologies linguistiques plus inclusives.

## Listes de publications

Le contenu de cette thèse s'appuie principalement sur les publications suivantes :

- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2023. Phylogeny-inspired soft prompts for data-to-text generation in low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–198, Nusa Dua, Bali. Association for Computational Linguistics.
- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2024. Generating from AMRs into high and low-resource languages using phylogenetic knowledge and hierarchical QLoRA training (HQL). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 70–81, Tokyo, Japan. Association for Computational Linguistics.
- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2025. Semantic Evaluation of Multilingual Data-to-Text Generation via NLI Fine-Tuning: precision, recall, and F1 scores. In *Findings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics. (*To appear*).

# Chapter 1

# Introduction

The quest to efficiently store, catalog, and retrieve information has been a constant throughout human history. This pursuit has fundamentally shaped how societies access, interpret, and apply knowledge across generations. From Kallimachos' *Pinakes* (Blum, 1991) at the Library of Alexandria in the 3rd century BCE through the Universal Decimal Classification bibliographic system (McIlwaine, 1997) from the 19th century to modern knowledge bases like Wikidata (Vrandečić and Krötzsch, 2014), humans have relied on structured data to store knowledge.

Structured representations, particularly graph-based data that presents information as networks of concepts or entities interconnected by their relations, offer substantial advantages over unstructured formats like plain text. The standardized nature of graphs contributes to greater consistency across different languages, reduces ambiguity, and improves clarity. These properties enhance their suitability for computational processing and large-scale analysis. Notable examples include Wikidata (Vrandečić and Krötzsch, 2014) and Google's Knowledge Graph (Singhal, 2012), both of which significantly facilitate information retrieval. However, while graphs are ideal for machines, natural language remains more effective for communicating this information to humans in accessible and engaging ways. Studies have demonstrated that presenting information in natural language improves comprehension and decision-making compared to structured data alone (Gkatzia et al., 2016).

To benefit from both structured graph-based representation and unstructured natural language text, it is possible to process data in graph form and convert it into fluent natural language. This approach can be applied in different ways to multiple tasks, it can be used in the verbalization stage of Information Extraction (IE) systems (Koncel-Kedziorski et al., 2019), to boost question answering (QA) by enriching the context (Han and Gardent, 2023), or in machine translation (MT) by providing additional information to the model (Song et al., 2019). These tasks fall under the field of Natural Language Generation (NLG), a subset of Natural Language Processing (NLP) that focuses on producing human-readable text from diverse inputs. This thesis focuses on the Data-to-Text (D2T) generation task, particularly the Graph-to-Text (G2T) branch, where the input is a graph-structured representation and the output is natural language text.

As with any NLG task, G2T aims to generate fluent and grammatically correct natural language text. However, it also introduces specific challenges. Notably, the generated text must accurately reflect the content of the input graph, which requires both semantic precision (including only content from the graph) and semantic recall (including all content from the graph). These complementary goals define the broader concept of semantic faithfulness, which is central to evaluating G2T systems.

While progress has been made in G2T generation for several graph types, challenges persist in multilingual generation, particularly regarding data availability. Although there are thousands of languages in the world, and hundreds with over a million speakers, available resources tend to concentrate on a few (Ruder, 2022). For example, from the 65 million existing Wikipedia articles, a common source for G2T data, almost half of them belong to just 10 of the 342 supported languages (Wikimedia Foundation, 2025). Furthermore, human-written and validated (gold-quality) graph-text pairs are scarce even in English, given the time and expertise required to collect them. This lack of training and evaluation data directly impacts the performance of G2T systems across all languages, especially in low-resource (LR) ones.

To address these limitations, this thesis proposes methods that leverage phylogenetic information (linguistic knowledge about the relationships between languages) to guide multilingual G2T model training. By incorporating this information, the study aims to enhance performance in settings with limited data, using multilingual transfer learning for both efficiency and regularization, while mitigating the noise often associated with multilingual training.

Equally important is the challenge of evaluation. Most current G2T evaluation practices rely on reference-based metrics, which assess similarity between generated outputs and gold-standard texts. However, these metrics often fail to capture semantic faithfulness and are challenging to scale across languages that lack high-quality references. To overcome this, the thesis also explores referenceless evaluation approaches applicable to multiple languages.

## 1.1 Research Questions

The primary motivation behind this work is to enhance G2T generation and evaluation across multiple languages, with a particular focus on low-resource languages. Enhancing G2T in this context can help extend its communicative and informational benefits to a broader population, contributing to the democratization of language technologies.

To this end, this thesis focuses on two specific structured semantic representations commonly used in NLP: the Resource Description Framework (RDF) graphs, widely employed in knowledge bases and linked data applications, and the Abstract Meaning Representation (AMR) graphs, utilized in language processing tasks. Both formalisms offer structured and language-independent meaning representations, making them particularly suitable for multilingual G2T tasks.

Building on this context, the thesis examines how to improve text generation and evaluation from structured semantic representations in multilingual settings, with a special focus on low-resource settings. Focusing on RDF and AMR as input formats, it explores the impact of language phylogeny, model adaptation, and referenceless evaluation. The following research questions guide this investigation.

**RQ1. Can text generation from Resource Description Framework (RDF) graphs be improved in low-resource languages with limited training examples by fully fine-tuning a model with soft prompts enriched with phylogenetic information?**

G2T systems in low-resource languages often underperform due to the scarcity of training data. This question investigates whether combining multilingual knowledge and structured prompting can compensate for this scarcity in RDF-to-Text generation. The approach is intuitively motivated by linguistic proximity: related languages often share structural and lexical features, making it plausible for models to transfer knowledge from high-resource counterparts. By incorporating phylogenetically informed soft prompts into training, the model can potentially exploit cross-lingual similarities while minimizing transfer noise, thereby improving generation quality even with minimal supervision. The question focuses on three Celtic languages (Breton, Irish, and Welsh) where resources, while limited, are available.

**RQ2. Can text generation from Abstract Meaning Representation (AMR) graphs be improved using phylogenetic information to guide a model's training process in high- and low-resource languages?**

This question is also motivated by linguistic proximity, specifically exploring whether a hierarchical curriculum learning strategy structured around language family relationships might be a viable way to facilitate cross-lingual transfer while reducing training noise. This question expands the scope to a new type of input graph and twelve Indo-European languages: six HR languages (Dutch, English, French, German, Italian, Spanish) and six related LR languages (Limburgish, Tok Pisin, Haitian Creole, Luxembourgish, Sicilian, Asturian).

**RQ3. Can Natural Language Inference (NLI) be used as the base to develop a referenceless multilingual evaluation metric for multiple facets of semantic faithfulness in RDF-to-Text generation across high- and low-resource languages?**

The widespread use of reference-based metrics, coupled with the limited or outright lack of gold-quality references in most languages, complicates the evaluation of multilingual G2T systems. While referenceless metrics exist to address this dependence on gold-quality data, they are mostly English-centric and only provide a limited coverage of the complex evaluation process. This question inquires whether an NLI-based approach can assess semantic precision, recall, and F1 directly between RDF graphs and generated texts. The motivation behind it stems from independent advances in multilingual NLI and referenceless NLI-based metrics.

## 1.2 Thesis Outline

Following a foundational overview, the remainder of the thesis is structured around the three research questions described above, each explored in its dedicated chapter, and concluding with a synthesis chapter.

Chapter 2 (Background) lays the conceptual groundwork by surveying the evolution of Natural Language Generation (NLG), with emphasis on model architectures, training strategies, and adaptation techniques. It then narrows the scope to the Graph-to-Text (G2T) task, discussing the characteristics of input graphs, such as RDF and AMR, their corresponding datasets, modeling

approaches, and standard evaluation practices. This chapter situates the thesis within existing work and highlights current gaps in multilingual and low-resource scenarios.

Chapter 3 (RDF-to-Text Generation of Celtic Languages) addresses RQ1, investigating whether RDF-to-Text generation in low-resource languages can be improved using a structured soft prompting method informed by phylogenetic relationships. Focusing on Irish (Gl1), Welsh (Cym), and Breton (Bre), the chapter details the proposed Phylogeny-Inspired Task-Source-Target (PI-TST) prompt design, based on monolingual unsupervised pretraining and supervised fine-tuning. It also presents ablation and data-size experiments, analyzed through both automatic and human evaluations.

Chapter 4 (AMR-to-Text Generation of High- and Low-resource Languages) expands the multilingual scope to twelve Indo-European languages and shifts the input format from RDF to AMR, targeting RQ2. This chapter introduces the Hierarchical QLoRA (HQL) framework, a curriculum-based, multilingual training strategy that utilizes parameter-efficient fine-tuning and language family hierarchies. It systematically examines how phylogenetic relations and language grouping impact cross-lingual transfer, particularly in low-resource settings.

Chapter 5 (Referenceless Evaluation of Multilingual RDF-to-Text) responds to RQ3 by proposing a multilingual evaluation metric based on Natural Language Inference (NLI), designed to function without reference texts. This chapter details the construction of synthetic training data, the adaptation of an NLI model for regression-based scoring, and the development of a metric that quantifies semantic precision, recall, and F1. Correlation analyses against both traditional metrics and human judgments demonstrate its effectiveness across high- and low-resource languages.

Finally, Chapter 6 (Conclusion) consolidates the key findings of the thesis, reflecting on the strengths and limitations of the proposed methods. It outlines avenues for future research in multilingual G2T, particularly regarding linguistic inclusivity, model scalability, and evaluation interpretability.

## 1.3 List of Publications

The content of this thesis is primarily based on the following peer-reviewed publications:

- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2023. Phylogeny-inspired soft prompts for data-to-text generation in low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–198, Nusa Dua, Bali. Association for Computational Linguistics.
- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2024. Generating from AMRs into high and low-resource languages using phylogenetic knowledge and hierarchical QLoRA training (HQL). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 70–81, Tokyo, Japan. Association for Computational Linguistics.
- William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2025. Semantic Evaluation of Multilingual Data-to-Text Generation via NLI Fine-Tuning: Precision, Recall, and F1 Scores. In *Findings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics. (*To appear*).

# Chapter 2

# Background

**Contents**



This chapter presents a series of concepts, tools, and materials that help situate this thesis within the extensive body of existing literature. In particular, it covers some basics of current NLG technology and a more specific overview of different G2T tasks. The information presented in this chapter helps to understand better the specific research conducted in this thesis, which is explained in detail in Chapters 3, 4, and 5.

## 2.1 Basics of NLG

Natural Language Generation (NLG) is a field that studies the production of fluent and grammatically correct natural language text. This broad definition encompasses any input the system receives, including text, sound, images, video, tables, or graphs. It also allows for various tasks depending on the relationship between the system's input and the generated text, such as translation (Macdonald, 1954), story writing (Meehan, 1977), speech-to-text (Bahl et al., 1983), simplification (Chandrasekar et al., 1996), summarization (Knight and Marcu, 2000), paraphrasing (Barzilay and Lee, 2003), or image captioning (Mason and Charniak, 2014).

While every type of input and its associated tasks have unique characteristics and requirements, multiple techniques have become standard in the NLG domain because of their efficacy. The rest of this sub-chapter explores some of the most prominent methods currently in use.

### 2.1.1 Brief History

This section outlines relevant developments in the field of Natural Language Generation (NLG), tracing its evolution from the early symbolic approaches of the mid-20th century to the contemporary Neural methods that are the current standard.

Symbolic NLG approaches emerged in the mid-20th century, characterized by the use of explicit, handcrafted rules to manipulate symbols representing linguistic information. These systems operated on the premise that language generation could be achieved through logical inference and structured representations. Early applications included machine translation (MT), as demonstrated by the Georgetown-IBM experiment in 1954 (Macdonald, 1954), and conversational agents like ELIZA (Weizenbaum, 1966) and SHRDLU (Winograd, 1972). The incorporation of world knowledge (structured information about the world used to inform the system's output) enabled rudimentary inference capabilities, leading to advances in story understanding (Schank et al., 1973; Cullingford, 1979) and question answering (McKeown, 1982). By the late 1980s and early 1990s, linguistically motivated systems utilizing grammars (Kasper, 1989) and templates (Reiter et al., 1995) became prevalent. These systems typically employed a modular pipeline architecture, dividing the generation process into sub-tasks such as content determination, text structuring, sentence aggregation, lexicalization, referring expression generation, and linguistic realization (Reiter and Dale, 2000; Gatt and Krahmer, 2018). While this modularity allowed for fine-grained control and interpretability, it also introduced multiple points of failure. Errors in one module could propagate through the pipeline, leading to incoherent or ungrammatical outputs (Meteer, 1991; Robin and McKeown, 1996).

Starting in the late 1970s and gaining prominence in the early 1990s, statistical natural language generation (NLG) benefited from the increasing availability of computational power and larger corpora. Unlike symbolic approaches, statistical methods rely on probabilistic models to learn language patterns from data, reducing the need for manual rule creation. Hidden Markov Models (HMM), a stochastic technique to model systems, were initially applied to the Speech-to-Text task (Jelinek, 1976) and later substituted by n-gram models (Bahl et al., 1983). Similar n-gram models were also used for machine translation (Brown et al., 1990). These data-driven approaches extended to generating text from structured input like graphs (Langkilde and Knight, 1998a), numerical data files (Belz, 2005), and database records (Konstas and Lapata, 2013). While statistical models improved scalability and robustness, they often struggled with maintaining global coherence and fluency, as their reliance on local context limited their ability to capture long-range dependencies in language.

Neural NLG emerged in the early 2000s with the introduction of models based on neural networks. Bengio et al. (2003) proposed a Feedforward Neural Network (FNN) for language modeling, marking a shift towards distributed representations of words. Subsequent advancements included the use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks for sequence modeling. Sutskever et al. (2014) introduced the Sequence-to-Sequence (Seq2Seq) framework using LSTMs for machine translation. Around the same time, also for machine translation, Cho et al. (2014) proposed the RNN encoder-decoder architecture for machine translation, where an encoder processes the input sequence into a fixed-length vector, and a decoder generates the output sequence based on this vector. However, the fixed-length context vector was identified as a bottleneck, limiting the model's ability to handle long sequences. To address this, Bahdanau et al. (2014) introduced an attention mechanism, allowing the decoder to access all encoder states and focus on relevant parts of the input during generation.

Finally, Vaswani et al. (2017) further explored the concept of attention in their seminal paper *Attention is All You Need*, which proposed the transformer architecture: a model relying solely on attention mechanisms, eliminating recurrence. The transformer architecture has since become the foundation for state-of-the-art NLG systems, including the ones used in this research. Given its prominence, the following section offers a more detailed examination of the transformer architecture.

### 2.1.2 Transformers

The transformer architecture, shown in Figure 2.1, was introduced by Vaswani et al. (2017). It was developed at Google Brain and Google Research and was initially proposed for machine translation. Since then, it has become the standard for many Natural Language Processing and Natural Language Generation tasks. This section summarizes its core innovations, as it is the primary architecture employed in this thesis.

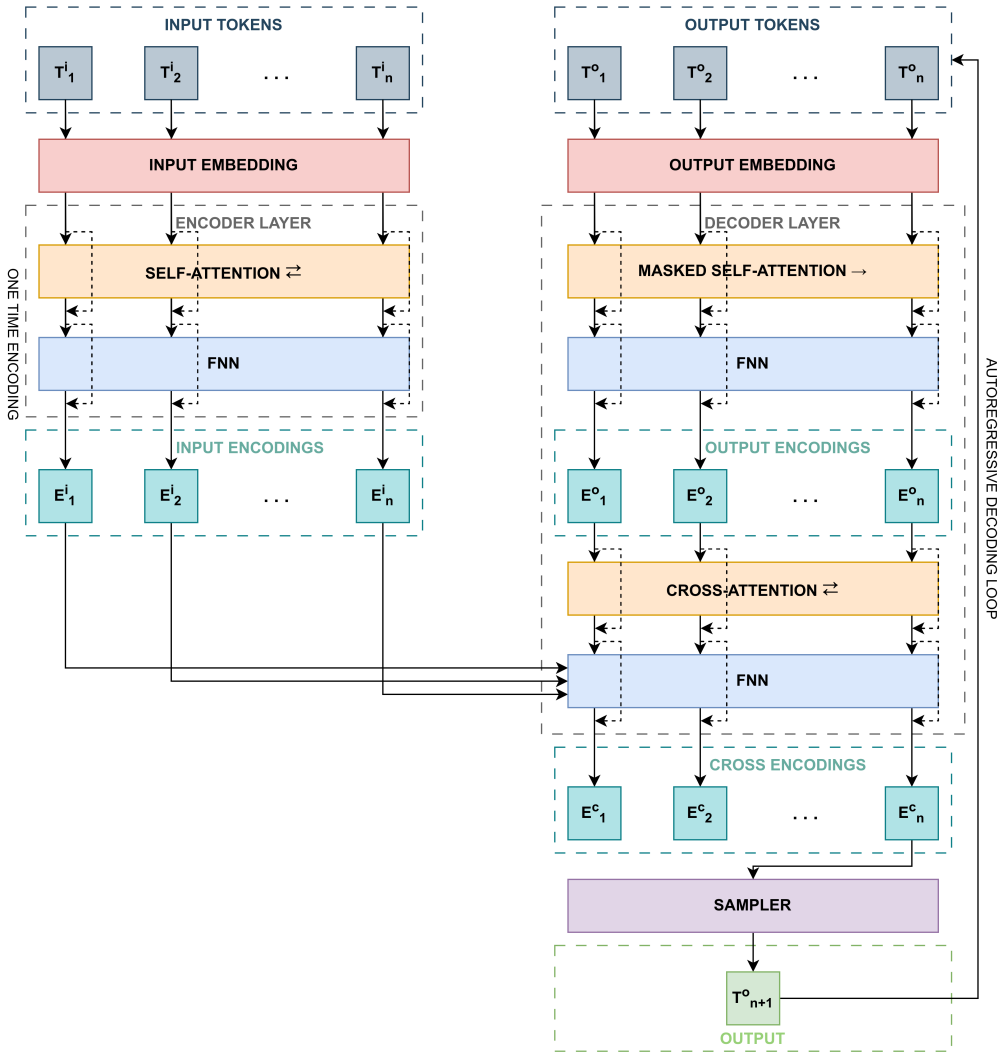



FIG. 2.1: Encoder-decoder transformer architecture simplified. Positional encodings and normalization layers are not shown for simplicity.

The main contribution of Vaswani et al. (2017) was the introduction of a novel encoder-decoder architecture: the transformer. This architecture relies solely on attention mechanisms, eliminating the need for recurrence or convolutions. The traditional Bidirectional RNN (BiRNN) encoder is replaced by a transformer encoder composed of a multi-head self-attention mechanism, which computes contextualized representations by attending to all positions in the input sequence. The self-attention is followed by a position-wise Feedforward Neural Network (FNN). Similarly, the Autoregressive RNN decoder is replaced by a transformer decoder, which includes a masked multi-head self-attention mechanism that ensures each position only attends to earlier positions in the output sequence, enforcing the autoregressive behavior. This self-attention is followed by a cross-attention mechanism, where the decoder attends to the encoder's output representations, and another position-wise FNN. Each sub-layer in both the encoder and decoder is wrapped with residual connections and followed by layer normalization, which stabilizes training. These innovations eliminate the sequential processing constraints of RNNs, enabling parallel processing during training and thereby improving efficiency. Moreover, the cross-attention mechanism enables the decoder to access all encoder outputs at each decoding step, rather than relying solely on the final hidden state as in traditional RNN architectures. This design facilitates the stacking of multiple transformer blocks, resulting in deeper models with increased capacity.

The architecture is centered around the Scaled Dot-Product attention mechanism, shown in Figure 2.2. This attention mechanism improved upon previous attention methods such as additive attention (Bahdanau et al., 2014) and multiplicative (dot-product) attention (Luong et al., 2015). The additive attention mechanism computes attention weights using an FNN that combines the decoder's current hidden state with each encoder hidden state. This approach is practical for variable-length input sequences and performs well at low dimension but suffers from a high computational cost due to the complexity of the FNN. A more efficient variant of the multiplicative attention mechanism was later proposed, based on the dot product between the decoder and encoder hidden states. Despite the gains in computational efficiency, it was consistently outperformed by the additive approach (Britz et al., 2017).

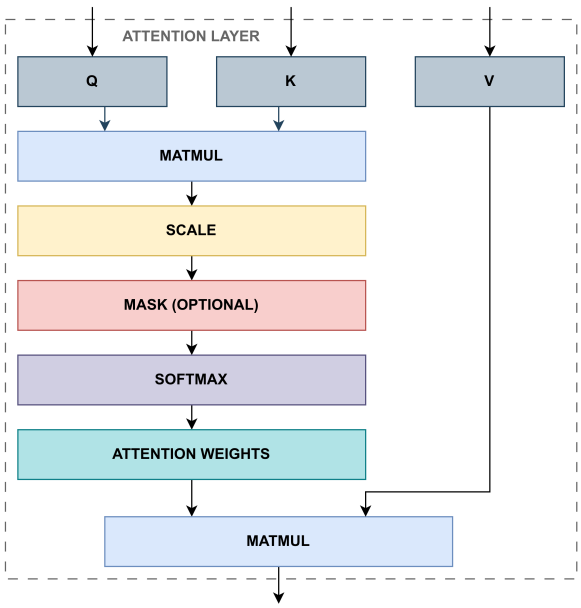



Fig. 2.2: Scaled dot-product attention mechanism.

Vaswani et al. (2017) noted that as vector dimensionality increases, the magnitude of their dot products also increases, potentially pushing the softmax to produce vanishing gradients. To address this, they introduced a scaling factor that divides the dot product by the square root of the input dimension before applying the softmax. Using that scaling maintains stable gradients during training and preserves the computational advantages of dot-product similarity without sacrificing performance. Furthermore, since the attention mechanism is parallelizable, all pairwise dot products can be computed as a matrix multiplication. This type of operation benefits from highly optimized implementations on modern hardware, particularly in Graphics Processing Units (GPUs).

A notable innovation of this attention mechanism is the use of separate projection matrices for Queries ($Q$), Keys ($K$), and Values ($V$). Prior approaches often used the same vectors for computing attention weights and transferring information. By using distinct linear projections, the transformer allows input sequences to be projected into multiple representation subspaces: $Q$ and $K$ are used to compute attention weights, and $V$ defines the transferred information.

The attention mechanism also supports an optional masking step, applied before the softmax operation, which sets certain positions to negative infinity, which turn to zero after the softmax. This masking prevents the model from attending to future tokens during training for autoregressive tasks such as language modeling, thus enabling the architecture to operate in either bidirectional or unidirectional modes.

In summary, if $Q$, $K$, and $V$ are matrices derived from the input embeddings through learned transformations, $M$ is the mask matrix, and $d_k$ is the dimension of the $K$ vectors, the scaled dot-product attention mechanism can be described by Equation 2.1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V \tag{2.1}$$

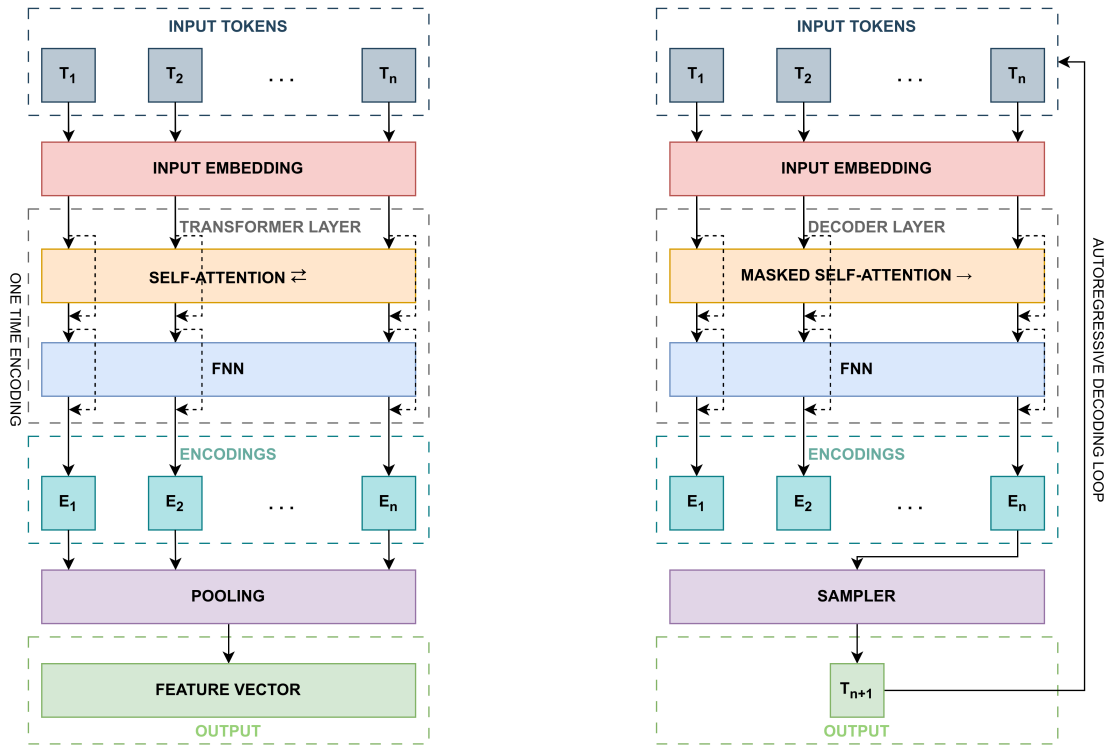
While the scaled dot-product attention mechanism enables parallel modeling of dependencies, a single attention head operates in only one representation subspace. Relying on a unique representation can limit the model's ability to capture diverse features across the input. To overcome this, the scaled dot-product attention of transformers has multi-head attention, which employs multiple attention heads, each with its own $Q$, $K$, and $V$ projections. These heads attend to information from different subspaces and positions in parallel. The outputs are concatenated and projected through another linear transformation to produce a final representation. In a setup with $h$ attention heads, where $W_i^Q$, $W_i^K$, and $W_i^V$ are the projection matrices for the $i$-th attention head, and $W^O$ is an output projection matrix, each attention head can be described by Equation 2.2, and the multi-head attention can be defined by Equation 2.3.

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.2}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.3}$$

At the start of both the encoder and decoder, token embeddings are passed through an embedding layer that maps them into a continuous vector space. As the model does not encode sequence order by default, a positional encoding is added to these embeddings. This encoding enables the model to capture both the relative and absolute positions of tokens within a sequence. Finally, the decoder output is passed through a linear transformation and softmax to generate a probability distribution over the vocabulary.

Although the transformer was initially presented in an encoder-decoder configuration, it quickly evolved into decoder-only and encoder-only variants. The decoder-only configuration, popularized by the Generative Pre-trained Transformer (GPT) from (Radford et al., 2018), uses only the decoder and omits the cross-attention mechanism. It employs masked self-attention and is optimized for autoregressive generation. Its streamlined design allows faster training, making it well-suited for large-scale language modeling. In contrast, the encoder-only configuration, famously started with the Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. (2019). This configuration retains only the encoder and has been highly successful for Natural Language Understanding (NLU) tasks, such as Natural Language Inference (NLI), Part-of-Speech (POS) tagging, sentiment analysis, and text embeddings. Figure 2.3 shows simplified diagrams of both configurations.




(a) Encoder-only

(b) Decoder-only

FIG. 2.3: Encoder-only and decoder-only transformer architectures simplified. Positional encodings and normalization layers are not shown for simplicity.

Despite the emergence of these variants, the encoder-decoder configuration remains central in NLG, especially in Sequence-to-Sequence (Seq2Seq) tasks, such as machine translation and Graph-to-Text, where full access to the input is essential. While the decoder-only models often excel in zero-shot and transfer learning settings, studies have shown that encoder-decoder models typically outperform them after supervised fine-tuning (Raffel et al., 2020; Wang et al., 2022; Zhang et al., 2022). Furthermore, encoder-decoder architectures offer efficiency advantages in low-parameter settings and are particularly well-suited for asymmetric tasks where input and output distributions differ (Elfeki et al., 2025).

### 2.1.3 Training Strategies

Natural Language Generation (NLG) training strategies have evolved significantly, encompassing various approaches to optimize model performance. This section covers the most predominant strategies, starting with basic early techniques, such as pre-training and fine-tuning, and progressing to more specialized concepts, including multilingual training and curriculum learning.

Initially, neural models for supervised tasks were trained from scratch, with random weight initialization. However, this often led to suboptimal solutions, especially in deep architectures, due to challenges like vanishing gradients and poor local minima. To address this, Hinton et al. (2006) introduced a greedy layer-wise unsupervised pre-training method for deep belief networks. This approach significantly improved convergence during subsequent supervised fine-tuning. Erhan et al. (2010) further demonstrated that unsupervised pre-training acts as a form of regularization, guiding the optimization process towards regions in the parameter space that support better generalization. In essence, pre-training enhances model performance by leveraging the abundance of unlabeled data, which is typically more accessible than labeled datasets.

Various transformation-based objectives have been employed to pre-train transformer models using unlabeled text, to instill general linguistic knowledge before task-specific fine-tuning. BERT (Devlin et al., 2019) utilized the Masked Language Modeling (MLM) objective, where a subset of input tokens is replaced with a [MASK] token, and the model learns to predict these masked tokens. GPT (Radford et al., 2018) and its successors (Radford et al., 2019; Brown et al., 2020) adopted the Causal Language Modeling (CLM) objective, training the model to predict the next token in a sequence, thereby enabling left-to-right text generation. The Text-to-Text Transfer Transformer (T5) by Raffel et al. (2020) introduced a span corruption objective, where contiguous spans of tokens are replaced with unique sentinel tokens, and the model is trained only to reconstruct the missing spans, facilitating a text-to-text framework. The Bidirectional and Auto-Regressive Transformer (BART) from Lewis et al. (2020) employs a denoising autoencoder approach. First, they applied a combination of noise functions like token masking, token deletion, text infilling, sentence permutation, and document rotation to corrupt the input text. Then, they pre-trained the model to reconstruct the original text from the corrupted version. All these objectives differ in their approach to capturing inherent language knowledge from unlabeled data. Table 2.1, on the next page, shows examples of these pre-training objectives.

| **Pre-training Objective** | **Input** | **Output** |
|---|---|---|
| Masked Language Modeling | The [MASK] sleeps. | cat |
| Causal Language Modeling | [BOS] The cat | The cat sleeps. |
| Span Corruption | The [X] sleeps. | [X] cat [Y] |
| Text Infilling/Masking† | [MASK] sleeps. | The cat sleeps. |
| Token Masking† | The [MASK] sleeps. | The cat sleeps. |
| Token Deletion† | The sleeps. | The cat sleeps. |
| Document Rotation† | sleeps. The cat | The cat sleeps. |

Tab. 2.1: Pre-training objectives examples applied to "The cat sleeps."
†These are some of the BART corruption objectives.

After pre-training to acquire general language knowledge, a model can be fine-tuned for specific tasks. Multiple versions of the same pre-trained model, such as BERT or GPT, can be fine-tuned on different datasets to specialize in various applications. In this case, all that is required is a labeled dataset consisting of specific inputs and associated outputs. Table 2.2 shows examples of some of those tasks.

| **Task** | **Input** | **Output** |
|---|---|---|
| Translation | The cat sleeps. | Le chat dort. |
| Text Classification | The cat sleeps. | Animals |
| POS Tagging | The cat sleeps. | DET NOUN VERB |
| Question Answering | The cat sleeps. Who sleeps? | The cat |
| Natural Language Inference | Cats are mammals. Cats are animals. | Entailment |

Tab. 2.2: Fine-tuning tasks examples.

Alternatively, fine-tuning can target multiple tasks simultaneously, as exemplified by the T5 model. This approach increases the amount of available supervision, which helps mitigate overfitting (where a model performs well on training data but poorly on unseen data), enhances generalization, and facilitates transfer learning (where knowledge learned from one task benefits another). However, fine-tuning carries the risk of catastrophic forgetting, particularly when moving from general-purpose to particular tasks.

There are two common strategies when fine-tuning on multiple tasks: 1) Supplementary Training on Intermediate Labeled data Tasks (STILTs), where a model is first fine-tuned on an auxiliary task before the main task, and 2) Multi-Task Learning (MTL), where several tasks are fine-tuned jointly. Weller et al. (2022) found that MTL performs better for target tasks with fewer training examples than the supporting task, whereas STILTs are more advantageous when the target task has more data than the supporting task.

Similar to training a single model to perform multiple tasks simultaneously, it is also possible to train a single model to process or generate various languages concurrently. This approach offers benefits, including increased efficiency by reducing the number of models required for diverse use cases and facilitating knowledge transfer across different languages. However, it is essential to acknowledge potential challenges, including negative transfer effects across languages or the risk of overfitting to high-resource languages if data sampling is not balanced correctly. Aharoni et al. (2019) introduced the first multilingual MT transformer.

Subsequent models like mBART (Liu et al., 2020) based on BARTS's denoising approach, M2M-100 (Fan et al., 2021) that uses language-specific sparse parameters, or NLLB (NLLB Team et al., 2022) which makes use of Mixture of Experts. Beyond MT, general-purpose multilingual pre-trained models, such as mT5 Xue et al. (2021), which uses T5's span corruption objective, have also been developed.

Finally, inspired by human learning processes, curriculum learning presents training data in a meaningful order rather than randomly. Bengio et al. (2009) proposed this approach, where the training data is ordered according to a predefined criterion (the curriculum) to guide the model toward better optima during training progressively. Similar to pre-training, applying an adequate curriculum provides a regularizing effect that enhances the model's generalization capabilities. In its original formulation, the curriculum started with texts containing frequent words and progressively introduced rarer vocabulary.

Subsequent work has explored alternative curricula. Xu et al. (2020) proposed a dynamic strategies that adjust sample difficulty based on training loss improvements. NLLB Team et al. (2022) adopted an approach that initially focuses on high-resource languages and gradually incorporates low-resource languages based on scarcity. Kuwanto et al. (2023) proposed a method that first trains on monolingual and code-switching data before introducing direct translation.

### 2.1.4 Model Adaptation

Since the early days of statistical NLG, researchers have been aware of the trade-offs when training on different data sources. While more data generally improves generalization, it often diminishes performance in domain-specific tasks. Model adaptation techniques emerged to address this challenge, aiming to optimize both data efficiency and computational resources while still enabling specialization across domains. These methods typically involve training a large base model on broad data and subsequently adapting it to specific domains or tasks without the need to retrain the entire model again.

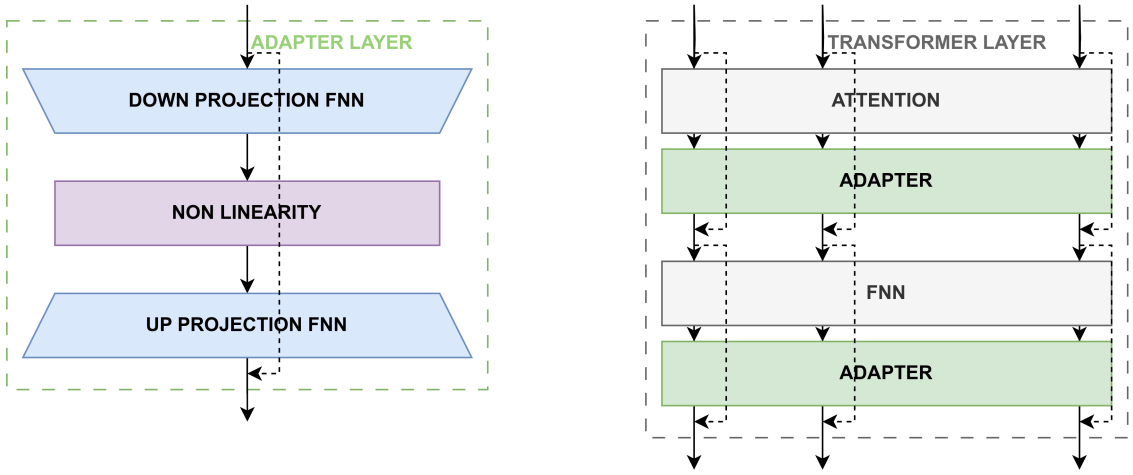
In pre-neural NLG systems, model adaptation techniques such as weight mixing were employed. These techniques involved interpolating model parameters from domain-specific and general-purpose models. A standard statistical approach was maximum a posteriori (MAP) adaptation, which updated model parameters by combining prior estimates with evidence from the target domain. This approach was successfully applied to text-to-speech (Bacchiani and Roark, 2003). Another family of techniques, discriminative training, directly optimizes decision boundaries using task-specific objectives rather than generative likelihoods. This approach was applied to tasks like Kana-to-Kanji conversion (Gao et al., 2006) and machine translation (Eidelman et al., 2012).

As neural models became mainstream, adaptation approaches evolved. Early methods involved fine-tuning only specific components of a model, such as the prediction heads, by freezing the remaining parameters. Similar approaches consisted of inserting additional trainable layers. These methods proved to be useful in text-to-speech (Ma et al., 2017). Another methodology used at the time was cost weighting, which emphasized in-domain examples during training and was successfully tried on machine translation (Chen et al., 2017).

With the rise of transformer-based architectures, which have led to a sharp increase in both parameter counts and data diversity requirements, the demand for more efficient adaptation techniques has intensified. Researchers began developing methods that reduced the number

of trainable parameters and enabled faster, modular adaptation without compromising model performance.

One prominent approach is the use of Adapter modules. Initially applied to visual (Rebuffi et al., 2017) and textual (Ma et al., 2017) neural models, adapters were later introduced into the transformer architecture by Houlsby et al. (2019). They proposed inserting lightweight bottleneck adapter layers after each transformer's attention and feedforward layers. During fine-tuning, only these adapter layers, the layer normalization parameters, and the prediction head were updated. The remaining weights remained unchanged during training (frozen). This approach enabled efficient transfer learning by preserving the general knowledge of the base model, thereby facilitating effective adaptation. The modular nature of adapters also facilitated dynamic switching between tasks, domains, or languages without requiring full fine-tuning. Though initially designed for encoder-only models, adapter modules were successfully extended to encoder-decoder setups in multilingual machine translation (Bapna and Firat, 2019). Figure 2.4 shows a simplified representation of a bottleneck adapter layer and its position inside a transformer layer.

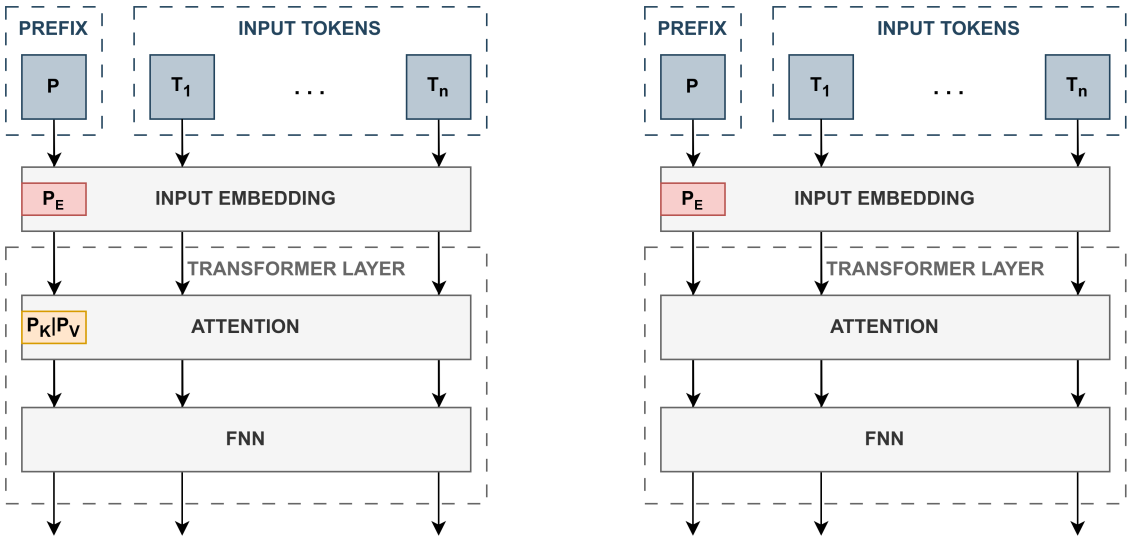



(a) Bottleneck adapter layer

(b) Transformer layer with adapters

FIG. 2.4: Bottleneck adapter simplified and its position inside a transformer layer. Normalization layers are not shown for simplicity.

As language models grew larger, it became possible to condition a frozen model toward specific tasks by crafting suitable prompts (Brown et al., 2020; Zhao et al., 2023). However, since finding optimal prompts is nontrivial, prompt-based adaptations gained traction. One such technique is Prefix-Tuning, introduced by Li and Liang (2021). The technique consists of prepending learnable vectors (called prefixes) to the input embeddings and the Key ($K$) and Value ($V$) matrices of the attention mechanism. This technique enables more precise control over the model's output while requiring only a small number of trainable parameters, thereby increasing memory efficiency and reducing the risk of overfitting. Building on this, Lester et al. (2021) proposed prompt-tuning, also called soft prompts: trainable vectors added only to the input embeddings, without modifying any other layer. This approach further enhanced parameter efficiency and modularity by isolating the adaptation to a single swappable component. Figure 2.5 shows these two methods, which only differ in the learnable weight of the attention layer.




(a) Prefix-Tuning

(b) Prompt-Tuning

FIG. 2.5: Prefix-Tuning and Prompt-Tuning simplified. Positional encodings and normalization layers are not shown for simplicity.

Despite their advantages, early adapter and prompt-based methods came with a trade-off. Adding new layers or vectors can sometimes increase inference latency or reduce the available input length. Moreover, these methods often underperformed when compared with full fine-tuning (FFT), requiring the finding of a balance between efficiency and task performance.

To mitigate these issues, Hu et al. (2022) proposed the Low-Rank Adapters(LoRA) architecture. Motivated by findings from Aghajanyan et al. (2021), they hypothesized that the weight updates in fine-tuning reside in a low-dimensional subspace. Consequently, they decomposed the weight update matrix into the product of two low-rank matrices and focused on only learning those matrices. In doing so, they were able to significantly reduce the number of trainable parameters while preserving and even outperforming full fine-tuning.

In transformer models, LoRA is typically applied to both attention and feedforward layers. During training, these low-rank matrices are learned separately and can later be merged into the original model weights. By merging the LoRA, the inference overhead of the additional parameters is eliminated without compromising performance quality. This modularity enables quick domain switching by swapping in different LoRA modules without affecting the rest of the model.

Mathematically, given an input vector $x$ and a pre-trained weights matrix $W_0$, the LoRA represents weight update as the product $\Delta W = BA$, where $B$ and $A$ are low-rank matrices of rank $r$ learned during fine-tuning. The update output is then represented by Equation 2.4.

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{2.4}$$

Alternatively, Figure 2.6 shows a schematic of the architecture.

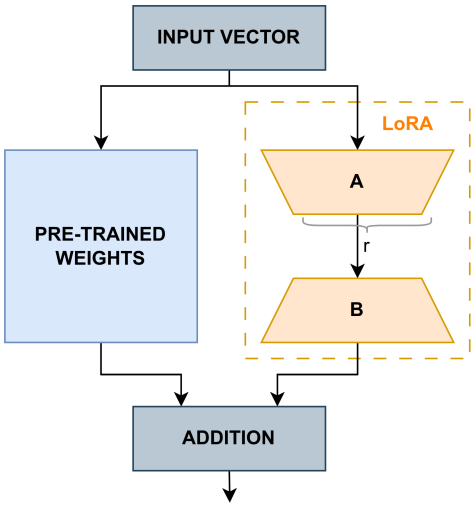



Fig. 2.6: Low-Rank Adapters (LoRA) simplified.

Finally, the growing size of transformer models made even LoRA fine-tuning increasingly complex. While quantization, the technique of using lower-precision representations to reduce memory usage, helped during inference (Hubara et al., 2018), it often caused instability during training. To overcome this, Dettmers et al. (2023) introduced Quantized LoRA (QLoRA), which performs low-rank fine-tuning on quantized base models while keeping the LoRA modules in full precision. This approach allowed efficient fine-tuning of very large models under limited resource constraints.

## 2.2 Graph-to-Text

Having covered several relevant NLG concepts, the following sub-chapter deals with the specific task addressed in this research: Graph-to-Text (G2T) generation.

As previously discussed, graph-structured data offers a clear and less ambiguous representation of information, facilitating both human understanding and computational processing. Moreover, graph structures promote cross-lingual consistency, as the same underlying graph can represent equivalent information across multiple languages. However, while graphs excel at structuring knowledge, natural language remains the most effective medium for communicating this information to humans in an accessible and engaging manner (Gkatzia et al., 2016). In this context, he aim of G2T generation is to combine the computational advantages of graph-based data with the communicative richness of natural language text.

This section provides an overview of the types of graphs commonly used as input for G2T systems, including existing datasets categorized by input type and language coverage, prevalent approaches to the task, and evaluation strategies. The nature of the input graph determines the structure and semantics that the model must handle; the dataset composition affects both training and generalization; and the evaluation metrics shape the understanding of how a model performs.

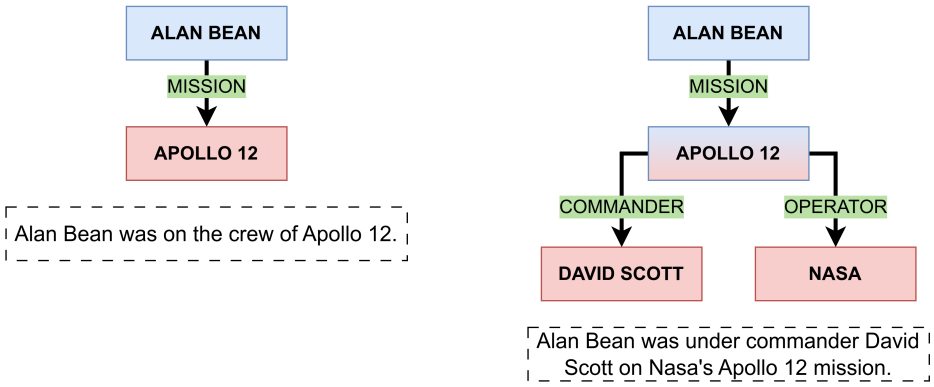
### 2.2.1 Types of Inputs

Just like there are many possible inputs to NLG systems, there are many possible inputs to G2T systems. This section introduces the two types of graphs studied during this research: Resource Description Framework (RDF) graphs and Abstract Meaning Representation (AMR) graphs. Additionally, a brief discussion is provided on other input formats that can be rendered as graph structures and have been investigated in the context of G2T before.

#### Resource Description Framework

According to the original RDF specification (Lassila and Swick, 1999), proposed by the World Wide Web Consortium (W3C), most data available on the web was machine-readable, but not machine-understandable. RDF was introduced to address this gap by enriching web data with metadata, enabling automatic systems to perform tasks such as resource discovery, cataloging, and content rating. It provided a standardized model for representing, encoding, and transferring such metadata through a graph-based structure.

Over time, graph structures proved helpful not only for metadata but for a wide variety of information types. As early as 2007, the DBpedia project (Auer et al., 2007) built an RDF knowledge graph from Wikipedia, with structured data extracted with relational databases and unstructured sources like infoboxes. Later, in 2012, Google introduced the concept of the Knowledge Graph to enhance search results with structured semantic content (Singhal, 2012). Following this trend, the Wikimedia Foundation launched Wikidata (Vrandečić and Krötzsch, 2014), a collaborative, multilingual, and constantly updated knowledge base. Wikidata adopted the RDF format (Erxleben et al., 2014) and, when possible, included links to DBpedia, thus reinforcing the importance of RDF in structured knowledge representation.




(a) RDF triple. (b) RDF graph.

FIG. 2.7: RDF triple and graph. Subjects are shown in blue, Predicates in green, and Objects in red. Some nodes can be simultaneously both the Subject and the Object.

As shown in Figure 2.7, graphs consist of a collection of RDF triples. Each triple contains a Subject node (the resource), a Predicate edge (the property or relationship), and an Object node (either a literal value or another resource). The Predicate forms a directed link from Subject to Object, resulting in labeled, directed graphs ideal for encoding complex knowledge in a machine-interpretable form. These properties make RDF a foundational input format in G2T systems, especially for tasks that require grounded, fact-based generation.

### Abstract Meaning Representation

An early version of AMR was introduced by Langkilde and KnightLangkilde and Knight (1998b), inspired by the Penman Sentence Plan Language (Kasper, 1989). These rooted, labeled, directed graphs encoded semantics where nodes denote concepts and edges encode relations. The formulation was built on the SENSUS knowledge base (Knight and Luk, 1994) and integrated lexical resources such as WordNet (Miller et al., 1990). At the time, the central goal was to create a unified semantic representation in which semantically equivalent sentences would yield identical graph structures, abstracting away syntactic variability.

This idea was later formalized by Banarescu et al. (2013), who based predicate representations on PropBank framesets (Kingsbury and Palmer, 2002), defined around 100 relation types, and introduced methods for encoding named entities and coreference. This standardization made AMR broadly applicable to both semantic parsing and generation tasks. Figure 2.8 illustrates a standardized AMR graph with multiple possible lexicalizations.

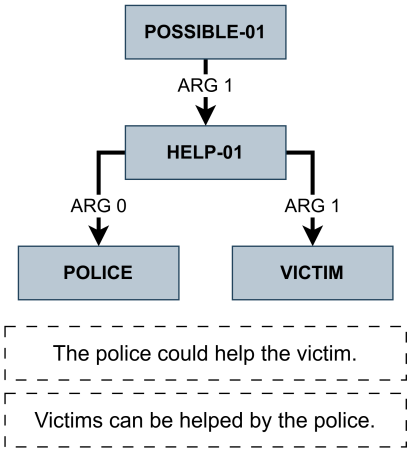



Fig. 2.8: AMR graph.

Recent variants, such as Minimal Recursion Semantics (MRS) by Hajdik et al. (2019), Uniform Meaning Representation (UMR) by Van Gysel et al. (2021), or BabelNet Meaning Representation (BMR) by Martínez Lorenzo et al. (2022), have been proposed to improve AMR's cross-linguistic applicability. However, the limited availability of high-quality annotated datasets and parsers for these variants has constrained their broader adoption.

#### Other Inputs

Besides RDF and AMR graphs, other structured data inputs have been explored in G2T generation. For instance, the Web Ontology Language (OWL) (Bechhofer et al., 2004) provides rich expressive constructs derived from Description Logic, enabling the representation of complex hierarchies and constraints. In addition, various knowledge graph structures do not follow a standardized framework (Koncel-Kedziorski et al., 2019). Similarly, entity-centric key–value paired datasets, where each record comprises attribute–value pairs for a single subject, serve as lightweight knowledge graph analogues (Novikova et al., 2017). These types of non-standard knowledge graphs can also be extracted from Table-to-Text data, usually after performing some content selection (Lebret et al., 2016).

Moving away from knowledge graphs, other graph structures like Universal Dependency trees, that display grammatical relations, can be used as the basis of syntax-driven G2T systems (Mille et al., 2017a). Finally, collections of predicate logic formulas, which model information as logical predicates and arguments, can also be represented as graphs (Chen and Mooney, 2008).

In summary, while RDF and AMR remain some of the major graph representations in G2T research due to their standardized format and data availability, a rich variety of alternative input structures continues to expand the field's boundaries.

### 2.2.2 Datasets and Languages

Transformers can perform across multiple tasks; however, like other deep neural networks, they require substantial training data to achieve optimal results. Moreover, gold standard datasets (those verified by humans) are generally required to evaluate generation quality (more on this in Subsection 2.2.4). As a result, the development of G2T systems is highly dependent on the availability and quality of datasets in each target language.

This section presents standard datasets used in RDF-to-Text, AMR-to-Text, and other related G2T tasks. Table 2.3, on the next page, summarizes some punctual information about these datasets. Afterwards, a more detailed description of each dataset is provided.

| Input Type | Dataset | Year | Languages† | Size‡ |
|---|---|---|---|---|
| RDF Graphs | WebNLG | 2017 | **Eng** | 25K |
| | | 2018 | Deu | 25K |
| | | 2020 | **Eng, Por, Rus** | 4K-45K |
| | | 2023 | **Bre, Cym, Gle, Mlt** | 3K |
| | | 2024 | **Xho** | 5K |
| | | 2025 | **Spa** | 45K |
| | WITA | 2020 | Eng | 55K |
| | KGTEXT | 2020 | Eng | 16M |
| | GenWiki | 2020 | Eng | 1.3M |
| | TekGen | 2021 | Eng | 57K |
| | KELM | 2021 | Eng | 15M |
| AMR Graphs | AMR | 2014 | **Eng** | 13K |
| | | 2017 | **Eng** | 39K |
| | | 2020 | **Eng, Deu, Ita, Spa, Zho** | 39K-59K |
| | TLP-AMR | 2014 | **Eng** | 1.5K |
| | | 2016 | **Zho** | 1.5K |
| | | 2024 | **Hrv, Kor** | 1.5K |
| | BIO-AMR | 2017 | **Eng** | 6K |
| | Europarl-AMR | 2020 | Bul, Ces, Dan, Deu, Ell, Eng, Est, Fin, Fra, Hun, Ita, Lav, Lit, Nld, Pol, Por, Ron, Slk, Slv, Spa, Swe | 400K-8M |
| | MASSIVE-AMR | 2024 | **Afr, Amh, Ara, Aze, Ben, Cym, Dan, Deu, Ell, Eng, Fas, Fin, Fra, Heb, Hin, Hun, Hye, Ind, Isl, Ita, Jpn, Jav, Kat, Khm, Kan, Kor, Lav, Mal, Mon, Msa, Mya, Nob, Nld, Pol, Por, Ron, Rus, Slv, Spa, Sqi, Swe, Swa, Tam, Tel, Tgl, Tha, Tur, Urd, Vie, Zho** | 1.6K |
| | MOSAICo | 2024 | Deu, Eng, Fra, Ita, Spa | 5M-17M |
| Non-RDF KGs | AGENDA | 2019 | Eng | 40K |
| | DART | 2021 | **Eng** | 82K |
| Key-Value Pairs | E2E | 2017 | **Eng** | 51K |
| | CACAPO | 2020 | **Eng, Nld** | 10k |
| Tables | WikiBio | 2016 | Eng | 728K |
| | RotoWire | 2017 | **Eng** | 5K |
| | WikiGen | 2018 | Eng | 200K |
| | ToTTo | 2020 | **Eng** | 136K |
| Logic Formulas | RoboCup | 2008 | **Eng** | 1.9K |
| | WEATHERGOV | 2009 | **Eng** | 22K |

Tab. 2.3: Available datasets by type of input. †Bold languages have gold-quality. ‡Approximate size per language, when it differs by language, the lower and upper bounds are presented.

### RDF Graphs

The 2017 *WebNLG* Challenge (Gardent et al., 2017) consisted of mapping sets of RDF triples to text. Along with the challenge, the largest existing gold-quality RDF-to-Text dataset was released. This dataset consisted of multiple RDF graphs, each paired with multiple English lexicalizations, similar to Figure 2.7b. The graphs were constructed from RDF triples extracted from DBPedia (Auer et al., 2007) across 15 distinct categories, five of which were exclusive to the test split and therefore unseen in the training split. The lexicalizations were obtained via crowdsourcing and then filtered and post-edited in the same way.

Following this, the 2020 *WebNLG* Challenge (Castro Ferreira et al., 2020)[1] extended the task to both G2T and semantic parsing in English and Russian. A new and unseen test split was created, and a new unseen category was added to it. Additionally, many lexicalizations were improved, and new information regarding the tree shape of the graphs was provided. For Russian, a subset of nine categories was selected, and the lexicalizations were translated using machine translation (Sennrich et al., 2017a). These translations were then post-edited via crowdsourcing Shimorina et al. (2019).

The 2023 *WebNLG* Challenge (Cripwell et al., 2023)[2] expanded support to under-resource languages. It introduced three low-resource (LR) Celtic languages (Breton, Irish, and Welsh) and Maltese. Professional translators translated the English development and test sets from 2020 into these languages. However, the training data was generated via machine translation (MT) and not post-edited, given the high costs of annotating low-resource (LR) languages.

Beyond the datasets associated with the official *WebNLG* Challenges, several versions of the dataset have been released by other authors. Castro Ferreira et al. (2018) created a German version[3] via MT but without post-editing. Almeida Costa et al. (2020) released a Portuguese version[4] of the test split using MT followed by human post-editing. Meyer and Buys (2024) translated all single-triple graphs into Xhosa[5] with native speaker assistance. Ramón-Ferrer et al. (2025) produced a full Spanish version[6] of the dataset, using MT and post-editing of low-quality translations.

Other RDF-based datasets exist; however, none of them is of gold quality since they rely on unsupervised alignment or synthetic generation. Furthermore, they are only available in English. Fu et al. (2020) created *WITA*[7], a partially aligned dataset. They extracted the first sentence of Wikipedia articles and aligned them with RDF triples from Wikidata. To find the triples, they used named entity recognition (NER) to extract entities and then retrieve RDF triples that contained them. Chen et al. (2020) created *KGTEXT*[8] by collecting Wikipedia sentence containing two or more hyperlinks. Instead of using NER to find entities, they used the hyperlinks to retrieve RDF triples from Wikidata. Jin et al. (2020) created *GenWiki*[9] by also collecting text from Wikipedia and using article titles and hyperlinks to find related RDF triples. However, they queried DBpedia instead of Wikidata.

[1] https://github.com/WebNLG/challenge-2020
[2] https://github.com/WebNLG/2023-Challenge
[3] https://github.com/ThiagoCF05/webnlg
[4] https://github.com/ThiagoCF05/webnlg-pt
[5] https://github.com/francois-meyer/t2x
[6] https://github.com/virginia-r99/Spanish_WebNLG_triples-to-text/
[7] https://github.com/fuzihaofzh/distant_supervision_nlg
[8] https://github.com/wenhuchen/KGPT
[9] https://github.com/zhijing-jin/genwiki

Finally, Agarwal et al. (2021) introduced two datasets: the Text from KG Generator(*TekGen*) dataset and the Knowledge Enhanced Language Model (*KELM*)[10] Pre-training corpus. *TekGen* pairs Wikipedia opening sentences with Wikidata RDF triples using distant supervision. In contrast, *KELM* is a large-scale corpus generated by a model trained on *TekGen* and fine-tuned on *WebNLG*.

### AMR Graphs

AMR datasets form another primary class of graph-to-text resources. The 2014 *AMR Annotation Release* (Knight et al., 2014)[11] initially consisted of English sentences from news wires, weblogs, and web discussion forums, parsed into AMR by professional annotators. A 2017 version (Knight et al., 2017)[12] expanded the dataset by incorporating data from broadcast conversations and doubled its size. Damonte and Cohen (2020)[13] later translated this version into Chinese, German, Italian, and Spanish. Finally, the 2020 version of the dataset (Knight et al., 2020)[14] further enlarged the dataset with new domains, including fiction and Wikipedia.

Besides the *AMR Annotation Release*, there are other AMR-Text datasets with both gold-quality text and graphs. In the literary domain, *TLP-AMR* (Banarescu et al., 2013) provides AMR annotations for every sentence from "The Little Prince". It was subsequently translated into Chinese (Li et al., 2016)[15], as well as Croatian and Korean (Kang et al., 2024)[16]. In the medical domain, *BIO-AMR* (May and Priyadarshi, 2017)[17] consists of AMR graphs for sentences from three full PubMed papers and 46 results sections. This resource remains English-only.

Finally, *MASSIVE-AMR* Regan et al. (2024)[18] took advantage of the multilingual parallel corpora of the MASSIVE dataset (FitzGerald et al., 2023)[19]. First, they tasked professional annotators to parse the 1 685 English sentences of MASSIVE into AMR graphs. Then, to expand to other languages, they paired the lexicalizations in different languages with the parsed graph and substituted its entity nodes with the corresponding language-specific version from MASSIVE. In this way, they obtained gold-quality pairs for 50 languages.

As with RDF, some AMR-Text datasets lack gold-quality, in this case due to their reliance on synthetic data. Fan and Gardent (2020) created an *Europarl-AMR* dataset[20] by automatically parsing into AMR the English portion of the Europarl dataset (Koehn, 2005)[21] using the JAMR parser (Flanigan et al., 2014). These synthetic AMRs were then paired with corresponding human translations in up to 21 European languages.

[10] https://github.com/google-research-datasets/KELM-corpus
[11] https://catalog.ldc.upenn.edu/LDC2014T12
[12] https://catalog.ldc.upenn.edu/LDC2017T10
[13] https://catalog.ldc.upenn.edu/LDC2020T07
[14] https://catalog.ldc.upenn.edu/LDC2020T02
[15] https://web.archive.org/web/20230602223634/https://amr.isi.edu/download/amr-bank-struct-v1.6.txt
[16] https://zenodo.org/records/14008284
[17] https://web.archive.org/web/20231207164411/https://amr.isi.edu/download/2018-01-25/amr-release-bio-v3.0.txt
[18] https://github.com/amazon-science/MASSIVE-AMR
[19] https://github.com/alexa/massive
[20] http://github.com/facebookresearch/m-amr2text
[21] https://www.statmt.org/europarl/

Finally, Conia et al. (2024) built *MOSAICo-AMR*[22] a multilingual AMR dataset from Wikipedia in five languages: English, French, German, Italian, and Spanish. They collected Wikipedia sentences and generated synthetic AMR graphs. For English, they used the LeakDistill parser (Vasylenko et al., 2023), and for the other languages, they used their implementation of the CLAP parser (Martinez Lorenzo and Navigli, 2024) trained on synthetic data.

### Non-RDF KGs

Some KG-to-Text datasets use non-standard representations instead of following defined frameworks, such as RDF. However, they are still structurally similar and can be interpreted as triples. The *AGENDA* dataset (Koncel-Kedziorski et al., 2019)[23] paired titles and abstracts of English scientific papers taken from the Semantic Scholar Corpus with knowledge graphs obtained by applying IE techniques on each scientific paper, particularly the SciIE system (Luan et al., 2018). *DART* (Nan et al., 2021)[24] was built from open-domain tables where entity-attribute-value triples were derived using a parent-child ontology, and humans generated lexicalizations for these extracted triples. Both of these datasets are only available in English.

### Key-Value Pairs

Some datasets use key-value lists as inputs, which can be transformed into triple-style inputs by extrapolating the entity. The *E2E* dataset (Novikova et al., 2017)[25] linked key-value restaurant descriptions with human-written English reviews. The *CACAPO* dataset van der Lee et al. (2020)[26] aligned key-value pairs with human news reports in domains such as weather, sports, stocks, and incidents. This dataset supports English and Dutch.

### Tables

Several datasets use structured tables as input, which can be transformed into triples after applying content selection. *RotoWire* (Wiseman et al., 2017)[27] collected professionally written, medium-length basketball game summaries and paired them with multiple tables of information about the game, including team and player statistics. *WikiBio* (Lebret et al., 2016)[28] paired the first sentence of Wikipedia Biography pages with the infobox table of the corresponding article. *WikiGen* (Perez-Beltrachini and Lapata, 2018)[29] expanded on WikiBio by using the entire first paragraph of the biographies instead of the first sentence. Additionally, they applied filtering to remove examples based on the number of properties in the infobox. *ToTTo* (Parikh et al., 2020)[30] first collected Wikipedia tables, excluding infoboxes, to avoid overlap with WikiBio and WikiGen. Then, selected sentences from the articles were retained that matched at least three table cells. All these datasets are only available in English.

[22] https://github.com/SapienzaNLP/mosaico
[23] https://github.com/rikdz/GraphWriter
[24] https://github.com/Yale-LILY/dart
[25] https://github.com/tuetschek/e2e-dataset
[26] https://github.com/TallChris91/CACAPO-Dataset
[27] https://github.com/harvardnlp/boxscore-data
[28] https://github.com/DavidGrangier/wikipedia-biography-dataset
[29] https://github.com/EdinburghNLP/wikigen
[30] https://github.com/google-research-datasets/ToTTo

#### Logic Formulas

Inputs in some datasets take the form of atomic predicate logic formulas. These can also be structured as triples. *RoboCup* (Chen and Mooney, 2008)[31] aligned soccer game commentary with logical event descriptions. *WEATHERGOV* (Liang et al., 2009)[32] paired weather forecasts with logical representations of meteorological data. Both are English-only.

### 2.2.3 Approaches

The following section covers some of the most relevant milestones in RDF-to-Text and AMR-to-Text, the specific G2T tasks studied during this thesis.

#### RDF-to-Text

Over the past decade, a wide range of methods have been proposed for the RDF-to-Text task, reflecting broader shifts in the field of natural language generation (NLG). These methods vary significantly in terms of architectural complexity, reliance on training data, and degree of linguistic control. This subsection categorizes and reviews the primary families of approaches that have influenced the development of RDF-to-Text generation, ranging from early rule-based systems to contemporary prompting techniques utilizing large language models. This organization not only highlights key milestones but also reveals underlying trends in the evolution of G2T modeling.

##### *Symbolic Approaches*

As discussed in Section 2.1, Symbolic NLG systems rely on handcrafted rules, templates, and structured linguistic resources. Below are some noteworthy symbolic systems used for RDF-to-Text generation.

In 2017, several symbolic approaches were proposed. *UITVNU-HCM* extracted rules from the typed dependency structure of the training text, allowing more syntactically informed rule creation. At generation time, WordNet (Miller et al., 1990) was used to compute predicate similarity and guide rule selection. *UTILBURG-PIPELINE* applied delexicalization to both triples and texts, extracted rules mapping triple structures to delexicalized outputs, and then generated text by matching input structures to rules. Then, a referring expression generation module (Castro Ferreira et al., 2016) filled in missing entities. *UPF-FORGE* employed handcrafted predicate-argument templates, which were then realized using the graph transducer FORGe system (Mille et al., 2017b).

In 2020, more symbolic systems emerged. *RALI* (Lapalme, 2020b) proposed a system that first grouped input triples into sentence-sized sets, mapped them to text using 200 manually defined templates, and produced a final output with the jsRealB surface realizer (Molins and Lapalme, 2015). *DANGNT-SGU* (Tran and Nguyen, 2020) used template extraction by replacing RDF subjects and objects with placeholders. During generation, templates were selected using the Jaro-Winkler similarity metric (Jaro, 1989) and applied to the input triples.

[31] https://www.cs.utexas.edu/~ml/clamp/sportscasting/#data
[32] https://link.zhihu.com/?target=https://cs.stanford.edu/~pliang/data/weather-data.zip

Despite the broader shift toward neural models, symbolic systems remain in use. *DCU/TCD-FORGe* (Mille et al., 2023) applied fully rule-based approach to RDF-to-Text generation into Irish. It was based on the graph transducer FORGe system (Mille et al., 2017b) and consisted of a 4-step pipeline: 1) triple lexicalization, 2) generation of non-inflected Irish text, 3) inflection generation, and 4) post-processing. *RDFpyrealb* (Lapalme, 2024) adopted a symbolic approach with a microplanning step using handmade rules followed by the use of the jsRealB surface realizer (Lapalme, 2020a) to generate the final text.

#### *Statistical Machine Translation (SMT)*

Before the rise of neural models, statistical approaches were the dominant method. *UTILBURG-SMT* utilized the Moses toolkit (Koehn et al., 2007) to train a model on a delexicalized version of the WebNLG 2017 dataset. Outputs were later relexicalized using alignment strategies and a 6-gram language model trained on Gigaword for ranking.

#### *Graph-aware LMs*

These models explicitly encode graph structure in their architecture to better capture the semantic relationships between components. *GTR-LSTM* (Trisedya et al., 2018) used a triple-level encoder followed by an LSTM decoder, where each RDF triple was encoded separately to preserve its internal structure. *DualEnc* (Zhao et al., 2020) used two Graph Convolutional Networks (GCN) (Kipf and Welling, 2017) as encoders: one to plan the content and the other to encode the information. Their outputs were combined and passed to an LSTM decoder to generate the final text. *Graformer* (Schmitt et al., 2021) enhanced the transformer architecture with a special encoder with graph-based attention layers. This encoder allowed the model to focus on graph topology during generation. *JointGT* (Ke et al., 2021) introduced a structure-aware semantic aggregation module. This module could be attached to the encoder layers of an existing PLM to help it preserve graph structure information. Additionally, they experimented with new pre-training objectives like graph-enhanced text reconstruction, text-enhanced graph reconstruction, and graph-text embedding alignment.

Wang et al. (2021) proposed a *Stage-wise* strategy that introduced two new position embeddings to the encoder of existing PLMs so they could learn information about the graph structure. One of the new embeddings encoded whether a sequence was a Subject, Predicate, or Object, while the other embedding encoded the position of a sequence within the graph structure.

#### *Fully Trained LMs*

Several systems trained encoder-decoder architectures end-to-end on linearized triple sequences, avoiding intermediate planning steps. *UTILBURG-NMT* adapted Edinburgh's WMT16 neural MT system (Sennrich et al., 2016), using delexicalized input-output pairs and a referring expression generator for post-processing. *ADAPTCENTRE* used the Nematus toolkit (Sennrich et al., 2017b) with byte-pair encoding subword tokenization and special tokens for triple separation.

*PKUWRITER* introduced a composite framework comprising a classic encoder-decoder architecture with attention mechanism, a ranker trained on synthetic data, and a reinforcement learning (RL) objective to improve content fidelity. The framework also used hand-crafted fallbacks to handle specific failure cases. *UMELBOURNE* enriched inputs by appending DBpedia entity types during delexicalization and used n-gram alignment to optimize target sequence matching. A standard attention-based encoder-decoder model generated outputs.

Moryossef et al. (2019) introduced *Step-by-Step*, a generation system that split RDF-to-Text into planning and realization. First, a content planner selected sentence structures; then a neural MT model (Gulcehre et al., 2016) generated the text. Castro Ferreira et al. (2019) proposed a more *Pipeline Transformer* approach leveraging neural models. While they experimented with multiple configurations, the best results were obtained with a 5-step pipeline of transformers: 1) discourse ordering, 2) text structuring, 3) lexicalization, 4) referring expression generation, and 5) textual realization.

*Huawei Noah's Ark Lab* (Zhou and Lampouras, 2020) used LASER embeddings (Schwenk and Douze, 2017) to guide guide template selection and delexicalization. *UPC-POE* (Bergés et al., 2020) generated additional silver-quality training pairs using back-translation (Domingo et al., 2020). Blinov (2020) proposed *med.* that started as a Russian GPT-2 model (Radford et al., 2019) later fine-tuned on Russian RDF-to-Text. They created silver-quality training data by machine translating the Chinese Baidu SKE dataset.

#### *Fine-tuned PLMs*

For a while, pre-trained models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) have become the backbone of most recent systems. In their *PLMs for G2T* study (Ribeiro et al., 2021b), they tried both approaches across several domains. Their approach consisted of first performing pre-training with related unlabeled corpora, followed by fine-tuning for different D2T tasks.

*NILC* (Sobrevilla Cabezudo and Pardo, 2020) fine-tuned BART directly on WebNLG 2020. *ORANGE-NLG* (Montella et al., 2020) applied noisy pretraining on RDF-text pairs extracted with the Stanford Open Information Extraction (Angeli et al., 2015) and used curriculum learning to improve robustness. *FBConvAI* (Yang et al., 2020) pre-trained BART on DocRED (Yao et al., 2019), a noisy parallel corpus sentences and automatically extracted relations. *HTLM:* (Aghajanyan et al., 2022) pre-trained BART on HTML documents to leverage the markup language's structured nature. They tested it on Zero- and One-shot experiments on multiple G2T datasets.

The *T2T Pre-Training for G2T* approach (Kale and Rastogi, 2020) tested different sizes of T5 on various G2T datasets *CycleGT* (Guo et al., 2020b) paired generation and parsing models in a back-translation loop to augment training data. *TGen* (Kertkeidkachorn and Takamura, 2020) ordered triples using position heuristics and then fine-tuned T5 for text generation. *NUIG-DSI* (Pasricha et al., 2020) pre-trained T5 on DBpedia abstracts before fine-tuning. *Amazon AI (Shanghai)* (Guo et al., 2020a) used a relational GCN for planning (Zhao et al., 2020) and a fine-tuned T5 model to lexicalize the plans. Clive et al. (2022) introduced *Control Prefixes* to steer generation via domain/task-specific embeddings. *Interno* (Kazakov et al., 2023) fine-tuned FRED-T5 on Russian RDF corpora, adding translation metadata to boost multilinguality.

The *cuni-ufal* model (Kasner and Dušek, 2020) was an mBART (Liu et al., 2020) fine-tuned on either English or Russian splits of the WebNLG dataset. *OSU Neural NLG* (Li et al., 2020) also fine-tuned mBART (Liu et al., 2020) on the Russian WebNLG and T5 (Raffel et al., 2020) on the English WebNLG. *bT5* (Agarwal et al., 2020) relied on multi-task fine-tuning using either T5 or mT5 (Xue et al., 2021) in a mixture of G2T, T2G, and MT data. *IREL* (Aditya Hari et al., 2023) and *CUNI-Wue* (Kumar et al., 2023) finetuned T5 to generate into English and then applied machine translation to the generated text.

*DCU-NLG-Small* Mille et al. (2024) consisted of a compact system with three main components: the FORGe system from Mille et al. (2023) for RDF-to-English, a T5 fine-tuned to paraphrase FORGe outputs into more fluent texts, and NLLB (NLLB Team et al., 2022) to translate the English outputs to other target languages.

*DipInfo-UniTo* Oliverio et al. (2024) introduce a pipeline approach with three steps: an algorithm to split the graph into subsets of no more than three triples, a Mistral (Jiang et al., 2023) or Llama-2 (Touvron et al., 2023) fine-tuned with QLoRAs (Dettmers et al., 2023) to lexicalize every subset individually, and the same LLM without fine-tuning to aggregate the independent texts. *DCU-ADAPT-modPB* Osuji et al. (2024) proposed a pipeline architecture using a fine-tuned Flan-T5 (Chung et al., 2024) for content ordering and content structuring, and Large Language Models (LLMS) like Mistral 7B (Jiang et al., 2023) for surface realization. The LLMs were either fine-tuned via LoRAs (Hu et al., 2022) or used zero-shot.

#### *LLM prompting*

These approaches involve various methods of utilizing LLMs off-the-shelf and without specialized fine-tuning for G2T.

For the 2023 WebNGL Challenge in low-resource languages, DCU-NLG-PBN (Lorandi and Belz, 2023) tried Zero-shot and Few-shot on GPT-3.5 to generate English text from an input graph. Afterwards, they translated the text to the target languages using the Google Translate API. Later, for the 2024 GEM Challenge, they applied a similar approach on Mistral 7B (Jiang et al., 2023) and Falcon-40B(Almazrouei et al., 2023). Additionally, they tried fine-tuning those models with LoRAs (Hu et al., 2022).

Also at the 2024 GEM Challenge, SaarLST (Jobanputra and Demberg, 2024) proposed using a symbolic retrieval system to find examples similar to the input graph. Then, they provide these examples as few-shot to Mixtral 8x7B (Jiang et al., 2024) or Command-R (Cohere, 2024) to generate the final text. Finally, OSU CompLing (Allen et al., 2024) tested three different approaches: Zero-shot GPT-4 (OpenAI et al., 2024) and fine-tuning Llama-2(Touvron et al., 2023) on synthetic data generated with GPT-4.

Table 2.4 summarizes some details about these approaches. It illustrates the evolution over time from symbolic approaches and models trained from scratch to fine-tuning and, more recently, prompting approaches. The table also highlights the strong bias towards English on RDF-to-Text models and the minimal number of systems tried in other languages.

| Type | Year | Approach | Languages |
|---|---|---|---|
| Symbolic | 2017 | UITVNU-HCM | Eng |
| | | UTILBURG-pipeline | Eng |
| | | UPF-Forge | Eng |
| | 2020 | RALI | Eng |
| | | DANGNT-SGU | Eng |
| | 2023 | DCU/TCD-Forge | Gle |
| | 2024 | RDFpyrelab | Eng |
| Statistical MT | 2017 | UTILBURG-SMT | Eng |
| Graph-aware LMs | 2018 | GTR-LSTM | Eng |
| | 2020 | DualEnc | Eng |
| | 2021 | Graformer | Eng |
| | | JointGT | Eng |
| | | Stage-wise | Eng |
| Trained LMs | 2017 | ADAPTCENTRE | Eng |
| | | UTILBURG-NMT | Eng |
| | | PKUWRITER | Eng |
| | | UMELBOURNE | Eng |
| | 2019 | Step-by-Step | Eng |
| | | Pipeline Transformer | Eng |
| | 2020 | UPC-POE | Eng |
| | | med. | Rus |
| | | Huawei Noahs Ark Lab | Eng, Rus |
| Fine-tuned PLMs | 2020 | T2T Pretraining for G2T | Eng |
| | | NILC | Eng |
| | | CycleGT | Eng |
| | | ORANGE-NLG | Eng |
| | | TGen | Eng |
| | | NUIG-DSI | Eng |
| | | Amazon AI | Eng |
| | | Cuni-ufal | Eng, Rus |
| | | FBConvAI | Eng, Rus |
| | | OSU Neural NLG | Eng, Rus |
| | | bT5 | Eng, Rus |
| | 2021 | PLMS for G2T | Eng |
| | 2022 | HTLM | Eng |
| | | Control Prefixes | Eng |
| | 2023 | Interno | Rus |
| | | IREL | Rus, Mlt, Gle, Cym |
| | | CUNI-Wue | Rus, Mlt, Gle, Cym, Bre |
| | 2024 | DipInfo-UniTo | Eng |
| | | DCU-NLG-Small | Ara, Deu, Eng, Hin, Kor, Rus, Spa, Swa, Zho |
| LLM prompting | 2024 | DCU-NLG-PBN(2023) | Mlt, Gle, Cym |
| | | SaarLST | Eng |
| | | OSU CompLing | Eng, Spa |
| | | DCU-Adapt-modPB | Eng, Hin, Kor, Swa |
| | | DCU-NLG-PBN(2024) | Ara, Deu, Eng, Hin, Kor, Rus, Spa, Swa, Zho |

Tab. 2.4: RDF-to-Text approaches by type and year.

### AMR-to-Text

Since their inception, even before being formalized, AMR graphs have been directly related to the NLG task (Langkilde and Knight, 1998b). However, for many years, progress in the field was hindered by the lack of a sufficiently large dataset. This situation changed with the 2014 AMR Annotation Release (Knight et al., 2014), which led to the emergence of multiple new approaches. This subsection categorizes major approaches to AMR-to-Text generation.

#### *Symbolic*

Flanigan et al. (2016) proposed a *Tree Transducer* system based on two main steps: generating an appropriate spanning tree from the AMR and then applying tree-to-string transducers to generate the final text.

#### *Statistical MT*

*Phrase-based MT*(Pourdamghani et al., 2016) proposed a pipeline with two steps: a linearization algorithm that arranged AMR nodes in English-like order, followed by a phrase-based MT system that turned linearization into natural text.

#### *Graph-aware LMs*

To better exploit the structure of AMR, several approaches incorporated specialized graph encoders. Song et al. (2016) proposed a pipeline involving AMR partitioning, local text generation, and sorting via the an *Asymmetric Generalized Traveling Salesman Problem* (AGTPS). In subsequent work, with the spread of neural models, Song et al. (2018) paired a graph encoder with an LSTM decoder and introduced a copying mechanism to address data sparsity. Beck et al. (2018) employed *Gated Graph Neural Networks* (GGNN) as encoders coupled with bidirectional RNN decoders. Guo et al. (2019) used *Densely Connected Graph Convolutional Network*(DCGCN) for their encoder paired with an LSTM decoder. Ribeiro et al. (2019) proposed *DualGraph* a model that had two different Graph Neural Networks Eecoders to capture different graph information (top-down and bottom-up). These representations were concatenated and given to a BiLSTM decoder. Finally, Bai et al. (2020) introduced an *Online Back-Parsing* approach using a graph encoder with a transformer decoder trained to generate sentences with embedded graph structure information.

#### *Trained LMs*

Castro Ferreira et al. (2017) explored the effects of three preprocessing steps on the input graph: delexicalization, compression, and linearization. They then trained both *Phrase and Neural MT* on the distinct preprocessed data. Konstas et al. (2017) also experimented with graph preprocessing in their *Neural AMR-to-Text*. In particular, the applied anonymization of named entities and linearizations was followed by training a neural machine translation (NMT) model to generate English sentences.

Other trained language models (LMs) focused on multilingual generation. Fan and Gardent (2020) trained a transformer encoder enriched with graph embeddings on a large synthetic AMR dataset based on *Europarl* and multiple transformer decoders trained on a large multilingual dataset. They then utilized cross-lingual sentence embeddings to connect the two elements. *XLPT-AMR* (Xu et al., 2021) was a transformer encoder-decoder model trained on a mixture of

real and synthetic data. The model was trained on a combination of tasks, including G2T, T2G, and MT for multiple languages.

#### *Fine-tuned PLMs*

Finally, as PLMs became widely available, some systems relied on them as the basis for fine-tuning. *GPT-Too* Mager et al. (2020) was a fine-tuned GPT-2 (Radford et al., 2019) where cycle consistency was reinforced by choosing the best generations during training. They train the model to reconstruct the graph before generating the output text. *DataTuner* Harkous et al. (2020) was also a fine-tuned GPT-2. It used a special semantic fidelity classifier to guide the sampling of the output text.

Also discussed in the RDF-to-Text section, the *PLMs for G2T* study by Ribeiro et al. (2021b) also experimented with AMR-to-Text. They further pre-trained T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) on text from similar domains before fine-tuning on task-specific datasets.

Ribeiro et al. (2021a) proposed the *Smelting Gold and Silver* approach, where they produced a large multilingual synthetic dataset leveraging AMR-parsing and MT. They performed multi-task fine-tuning on mT5 (Xue et al., 2021) using different combinations of their synthetic data.

*SPRING* Bevilacqua et al. (2021) tackled both G2T and T2G generation by finding the best linearization approach and fine-tuning BART (Lewis et al., 2020) on both tasks. A *Multilingual SPRING* model was trained on machine-translated data by Martínez Lorenzo et al. (2022).

*AMR-BART* Bai et al. (2022) also experimented with BART. They further pre-trained it for both G2T and T2G by giving the model a masked concatenation of linearized graph and text. In this way, the model learned to generate either the reconstructed graph or the reconstructed text. They fine-tuned it by masking only the tokens from the element to be generated. Finally, *BiBl* Cheng et al. (2022) was also a fine-tuned BART on G2T, T2G, and a reconstruction task where a masked concatenation of text and graph must be reconstructed entirely.

The development of AMR-to-Text generation has evolved significantly over the past decade, progressing from early symbolic systems to sophisticated, multilingual, fine-tuned models. Each paradigm reflects broader shifts in the field of Natural Language Generation, providing unique insights into the trade-offs between linguistic fidelity, scalability, and model complexity.

Table 2.5 below provides a concise overview of the major AMR-to-Text approaches categorized by type and year. The table illustrates the increasing trend towards leveraging pre-trained language models, while highlighting the scarcity of multilingual approaches.

| Type | Year | Approach | Languages |
|---|---|---|---|
| Symbolic | 2016 | Tree Transducer | Eng |
| Statistical MT | 2016 | Phrase-based MT | Eng |
| Graph-aware LMs | 2016 | AGTSP | Eng |
| | 2018 | GGNN | Eng |
| | 2019 | DCGCN | Eng |
| | | DualGraph | Eng |
| | 2020 | Online Back-Parsing | Eng |
| Trained LMs | 2017 | Phrase and Neural MT | Eng |
| | | Neural AMR-to-Text | Eng |
| | 2020 | Europarl | Bul, Ces, Dan, Deu, Ell, Eng, Est, Fin, Fra, Hun, Ita, Lav, Lit, Nld, Pol, Por, Ron, Slk, Slv, Spa, Swe |
| | 2021 | XLPT-AMR | Deu, Eng, Ita, Spa |
| Fine-tuned PLMs | 2020 | GPT-too | Eng |
| | | DataTuner | Eng |
| | 2021 | PLMs for G2T | Eng |
| | | SPRING | Eng |
| | | Smelting Gold and Silver | Deu, Ita, Spa, Zho |
| | 2022 | AMR-BART | Eng |
| | | BiBl | Eng |
| | | Multilingual SPRING | Deu, Eng, Ita, Spa |

Tab. 2.5: AMR-to-Text approaches by type and year.

### 2.2.4 Evaluation

A sound and reliable evaluation methodology is essential for correctly assessing the performance of any G2T system. Such methodologies ensure that the outputs of generation models are judged consistently and meaningfully across different datasets. Without a reliable evaluation framework, it becomes challenging to draw valid conclusions about model performance, resulting in misleading comparisons and ultimately, stagnation in progress due to a lack of actionable insights.

The following section examines some of the most common methods for evaluating G2T generation.

#### Human Evaluation

As discussed in Chapter 1, the goal of G2T is to produce text that can be easily communicated and understood by humans. Because of that, the best way to evaluate Natural Language Generation (NLG) is to test it on that exact task through performing human evaluations. However, if the instructions given to the evaluators are confusing or the descriptions reported in studies are unclear, the results of human evaluation might be unreliable (Belz et al., 2020; Howcroft et al., 2020). A possible way to address this problem is by using an established taxonomy for annotations, like the Quality Criteria for Evaluation of Text (Belz et al., 2024), or QCET, which describes in detail each type of evaluation. Below is a small and non-exhaustive selection of human evaluations generally used when evaluating G2T generations, along with their corresponding QCET taxonomy code.

Some metrics can be assessed by an evaluator simply by examining the generated text alone. These metrics generally provide an overall understanding of the quality of the language generated. Some of the most relevant metrics of this type are *Grammaticality (TQCO-f-1)* which refers to the degree to which a text is free of grammatical errors, *Fluency (QGO-b-2)* which describes the degree to which a text flows well, and *Readability (QGO-b-1)* which indicates the degree to which an output is easy to read.

Other metrics contrast the generated text with the provided input and generally provide information about the success with which the task was performed (besides language quality). For this research, the most relevant is *Correctness of outputs relative to input content (QCI-c)*, which describes the degree to which the content of a text is correct relative to the input and is closely related to semantic Faithfulness. This metric in turn can be split into two more specific metrics: *Absence of Omissions (QCI-c-1)* which meassures the degree to which the content of a text expresses the content of the input (related to semantic recall), and *Absence of Additions (QCI-c-2)* which is the degree to which the content of a text expresses only content present in the input (related to semantic precision).

Table 2.6 shows examples of good and bad text according to all the discussed metrics.

| **Input Triples**<br>Alan Bean, mission, Apollo 12<br>Apollo 12, commander, David Scott<br>Apollo 12, operator, NASA | |
|---|---|
| **God Generation**<br>Alan Bean was under commander David Scott on NASA's Apollo 12 mission. | |
| **Metric** | **Bad Example** |
| Grammaticality (TQCO-f-1) | Alan Bean were under commander david scott on nasas Apollo 12 mission |
| Fluency (QGO-b-2) | Alan Beam was on the crew of the Apollo 12. The Apollo 12 was under commander David Scott. The Apollo 12 was operated by NASA. |
| Readability (QGO-b-1) | NASA's mission Apollo 12, under commander David Scott, had as a member of its crew Alan Bean . |
| Content Correctness (QCI-c) | Alan Bean, born on March 15, 1932, was on NASA's Apollo 12 mission. |
| Absence of Omissions (QCI-c-1) | Alan Bean was on NASA's Apollo 12 mission. |
| Absence of Additions (QCI-c-2) | Alan Bean, born on March 15, 1932, was under Commander David Scott on NASA's Apollo 12 mission. |

Tab. 2.6: Human evaluation examples of good and bad text according to different metrics.

In addition to properly defining the characteristics to be evaluated, it is essential to define an evaluation approach that is both informative and easy to understand by the evaluators; a standard way of doing this is by using a Likert scale (Likert, 1932). Striking a good balance between ease of use and level of information collected is crucial: highly detailed scales (e.g., 0-100) might produce very fine-grained information but confuse the evaluator or make their work more tedious, while a scale too simple (e.g., 1-3) might not provide enough information to distinguish the generations properly.

Once the evaluation task has been defined, evaluators need to receive adequate training and vetting to ensure they will provide reliable evaluations. However, despite all the measures taken, evaluators are likely to introduce their subjective biases into the evaluation process. Because of this, it is recommended that several evaluators be used to corroborate the same generated texts. Afterward, the inter-annotator agreement can be evaluated with different metrics like Cohen's $\kappa$ (Cohen, 1960), Fleiss' $\kappa$ (Fleiss, 1971), or Krippendorff's $\alpha$ (Hayes and Krippendorff, 2007).

Ethical considerations must also be assessed beyond the technical and scientific aspects of the human evaluation process. These include factors like fair wages, privacy, and anonymity, or psychological risks (Shmueli et al., 2021).

### Automatic Evaluation

While a detailed and robust human evaluation with multiple evaluators provides a generation with a more informative and accurate evaluation, it can be expensive and time-consuming, if not outright impossible. Multiple automatic metrics have emerged over time to mitigate this obstacle. Below is a non-exhaustive list of some of the most common metrics used in G2T generation, many of which were initially proposed to evaluate machine translation (MT).

#### *Surface-Based Reference-Based Metrics*

The first metrics used to evaluate G2T generation compare the generation against one or multiple references based on their surface representation.

Bilingual Evaluation Understudy (Papineni et al., 2002), better known as *BLEU*, was proposed as a fast and effective way of evaluating *MT systems*. The metric primarily relies on token $n$-gram precision, which is the ratio of overlapping $n$-grams to the total number of $n$-grams in the candidate. Equation 2.5 shows how to compute the $n$-gram precision of a corpus for a given $n$, where $\text{Count}_{\text{Clip}}$ is clipped by the maximum count of a given $n$-gram across all the references of candidate $C$.

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{Clip}}(n\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram'} \in C'} \text{Count}(n\text{-gram'})} \tag{2.5}$$

A final BLEU-N score can be computed, as shown in Equation 2.6, by obtaining the weighted geometric average of all $n$-gram precisions $p_n$ from $n = 1$ until $n = N$ and multiplying that by a corpus-level Brevity Penalty (BP) used to punish under-generated outputs.

$$\text{BLEU-N} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.6}$$

The brevity penalty is computed at the corpus level following Equation 2.7, where $c$ is the summed length of all the candidates in the corpus and $r$ is the summed length of the corpus references, closer to the length of their respective candidates.

$$\text{BP} = \begin{cases} 1 \text{ if } c > r \\ e^{1-(r/c)} \text{ if } c \leq r \end{cases} \tag{2.7}$$

Thanks to its speed and ease of use, BLEU quickly became a staple in evaluating many NLG tasks. Despite its popularity, it has drawn criticism, including its use in tasks it was not designed for (Reiter, 2018), the impact of the tokenizer used, and the lack of clarity in reporting results (Post, 2018).

Furthermore, since it was designed to be a corpus measure, BLEU can have some undesirable properties when used for single sentences. To address this issue, Wu et al. (2016a) from Google proposed *GLEU*, which is the minimum score between precision and recall of tokens. According to their experiments, it correlates well with BLEU on a corpus level but has its drawbacks when evaluating individual sentences.

*METEOR* (Banerjee and Lavie, 2005) was another approach to mitigate some of the shortcomings of BLEU. Instead of relying solely on strict string matching to pair n-grams from the reference and the generation, METEOR allowed for other matching strategies, such as stemmed forms and meaning representations.

Through a series of steps utilizing these different matching mechanisms, METEOR identifies the largest one-to-one unigram alignment between the reference and the generation. Then, precision and recall are computed based on that alignment, and a weighted F score is produced with emphasis on recall, as shown by Equation 2.8.

$$\text{Fmean} = \frac{10PR}{R + 9P} \tag{2.8}$$

To account for word order and n-grams of n bigger than one, the authors group the matched uni-grams in the smallest possible number of chunks so that, within each chunk, uni-grams have the same order on both reference and generation. With this information, they compute a penalty score as per Equation 2.9.

$$\text{Penalty} = o.5 \cdot \left( \frac{\#\text{ chunks}}{\#matcheduni - grams} \right)^3 \tag{2.9}$$

The final score is then computed following Equation 2.10.

$$\text{METEOR} = \text{Fmean} \cdot (1 - \text{Penalty}) \tag{2.10}$$

Translation Edit Rate (Snover et al., 2006), also known as *TER*, takes a different approach than BLEU. In this case, the score is based on the number of required edits that need to be applied to a candidate generation so it matches its closest references, as shown in Equation 2.11. The list of possible edits includes insertion, deletion, and substitution of single words and shifts of word sequences (a shift moves a contiguous sequence of words within the candidate to a different position).

$$\text{TER} = \frac{\#\text{ of edits tot he closest reference}}{\text{average }\#\text{ of reference words}} \tag{2.11}$$

Character $n$-gram F-score (Popović, 2015), or *ChrF*, works with character $n$-grams. By acting at the character level, this metric disentangles itself from tokenization's impact, making it particularly useful for multilingual evaluation (especially in LR languages). The score is computed following Equation 2.12, where ChrP is the character $n$-gram precision (percentage of $n$-grams in the candidate that has a match in the reference), ChrR is the character $n$-gram recall (percentage of $n$-grams in the reference that have a match in the candidate), and $\beta$ is a parameter which assigns $\beta$ times more importance to recall than to precision.

$$\text{ChrF}\beta = (1 + \beta^2)\frac{\text{ChrP} \cdot \text{ChrR}}{(\beta^2 \cdot \text{ChrP}) + \text{ChrR}} \tag{2.12}$$

Further research on this metric (Popović, 2017) suggested $\beta = 2$ as the best parametrization. That research also discovered that enriching the score by counting word unigrams (ChrF+) and bigrams (ChrF++) also improved its performance.

#### *Model-Based Reference-Based Metrics*

With the emergence of the transformer architecture and its strong capabilities for NLU, new automatic metrics have been developed that rely on this new technology.

Reimers and Gurevych (2019) proposed Sentence-BERT, better known as *SBERT*, to compute Semantic Textual Similarity (STS) efficiently. This approach consisted of a modified training strategy for the BERT (Devlin et al., 2019) architecture. While BERT excels at various natural language processing tasks, it processes sentence pairs jointly, which can be computationally intensive for tasks like semantic similarity search and clustering.

To address this, SBERT proposed the use of a Siamese network architecture, where two identical BERT models with shared weights encoded sentences independently. Each sentence was then transformed into fixed-size vector representations through a pooling operation. Then, the similarity between sentences was computed using cosine similarity between their embeddings. This approach enabled rapid comparison of sentence embeddings, significantly reducing computational overhead compared to traditional BERT models. Figure 2.9 shows a simplified version of the configuration.

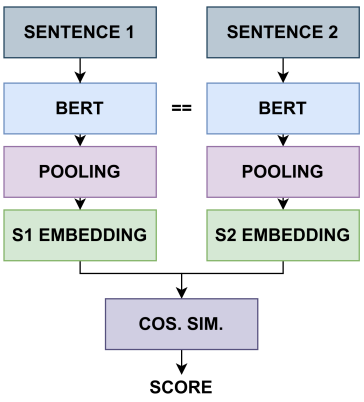



Fig. 2.9: Sentence-BERT (SBERT)simplified representation.

This type of Semantic Textual Similarity can be used to evaluate how closely the meaning of a candidate sentence aligns with its reference by obtaining the embeddings of both and measuring their cosine similarity. However, this requires an encoding model capable of generating sentence embeddings in the language being evaluated. For example, Language-agnostic BERT Sentence Embeddings (Feng et al., 2022), or *LABSE*, obtained with a model pre-trained and fine-tuned to measure the STS on texts across more than 100 languages.

Taking advantage of the contextual token embeddings from BERT, Zhang et al. (2020b) proposed *BERTScore*. With this approach, instead of pooling token embeddings into a unique sentence embedding and computing similarity at that scale, the authors computed a pairwise cosine similarity at the token level. They then computed precision, recall, and F scores by taking into consideration the pairs with the highest similarity. Equations 2.13, 2.14, and 2.15, show those computations, where $C$ and $R$ are the lists of pre-normalized contextual token embeddings of the candidate and reference texts, respectively, and $x^\top y$ is the inner product of two token embeddings.

$$\mathrm{P}_{BERT} = \frac{1}{|C|} \sum_{c_i \in C} \max_{r_j \in R} c_i^\top r_j \tag{2.13}$$

$$\mathrm{R}_{BERT} = \frac{1}{|R|} \sum_{r_i \in R} \max_{c_j \in c} r_i^\top c_j \tag{2.14}$$

$$\mathrm{F}_{BERT} = 2\frac{\mathrm{P}_{\mathrm{BERT}} \cdot \mathrm{R}_{\mathrm{BERT}}}{\mathrm{P}_{\mathrm{BERT}} + \mathrm{R}_{\mathrm{BERT}}} \tag{2.15}$$

Sellam et al. (2020) proposed *BLEURT* by instead focusing on the regression capabilities of BERT when used as a cross-encoder. This approach means that, given two concatenated sentences (reference and generation), the model can directly output a score. Instead of relying on cosine similarity, precision, or recall, BLEURT aimed to directly learn a scoring policy to differentiate good generations from bad ones.

The model was trained on three steps: 1) The standard BERT pre-training, 2) Further pre-training on synthetic reference-generation score pairs, where the generations are obtained by corrupting correct texts and the scores are obtained with automatic metrics like BLEU or BERTScore, and 3) Fine-tuning the model on real reference-generation score pairs, where the generations are real systems submissions to WMT and WebNLG Challenges and the scores are the public results of human evaluations. Later, Pu et al. (2021) introduced BLEURT-20, expanding the support to up to 20 languages.

#### *Model-Based Referenceless Metrics*

All the automatic metrics discussed until now require at least one reference to compare candidate generations against them. While they have all been proven to be valuable metrics in developing and assessing new NLG models, their reliance on gold-quality references can be an obstacle given the cost of producing such references, particularly in multilingual settings and when dealing with low-resource languages. Referenceless metrics that evaluate the generation independently or in comparison to the input text have been proposed to overcome this limitation.

Lau et al. (2017) found that the Syntactic Log-Odds Ratio (Pauls and Klein, 2012) or SLOR correlates well with acceptability judgments at the sentence level. The SLOR is a normalized log-probability score that adjusts a sentence's likelihood by accounting for its length and the frequencies of its words. Later, Kann et al. (2018) empirically proved that it can be used as an automatic proxy for fluency, and proposed a WordPiece variation (WPSLOR), which works with a smaller model.

To evaluate AMR-to-Text generation, Manning and Schneider (2021) proposed a *parsing-based* referenceless approach. It consisted of parsing the generated text back into AMR format and computing the similarity between the input AMR and the parsed output. They found that AMR parsers are still too noisy to perform the task correctly, but performance improves significantly when manual annotations are used.

To evaluate the semantic accuracy in RDF-to-Text generation, Dušek and Kasner (2020) proposed an off-the-shelf *NLI approach*. They used existing NLI models to compare generated texts against input graphs. First, they converted the graph's triples into natural language text by applying a template. Then, to check for omissions, the generated text was used as the premise, and each triple was individually tested as a hypothesis. If a given test was not marked as an entailment, the triple was considered omitted in the candidate text. Similarly, to check for additions, the candidate was used as the premise, and the entire graph as the hypothesis. If the test is not marked as an entailment, an addition is detected.

Zhang et al. (2023) proposed *FactSpotter*, which also relies on NLI models to compare generated texts with input graphs. However, this approach only focused on spotting omissions in the generated text. They followed a similar approach to Dušek and Kasner (2020), using the generation as the premise and the individual triples as the hypothesis. The main difference is that they did not reformat the triples; instead, they fine-tuned their model on synthetic data.

Finally, Le Scao and Gardent (2023) proposed *EREDAT*. This approach drew inspiration from SBERT but addressed the issue of reference scarcity. Instead of encoding a reference and a generation, they trained a model on graph-text pairs so they could directly compare the embeddings of generations and input graphs.

## 2.3 Conclusion

The current chapter provided the foundational concepts and background necessary to frame this thesis. Essential advancements in the general natural language generation (NLG) domain were reviewed, with an emphasis on the transformer architecture, its training, and related model adaptations. An overview of the Graph-to-Text task was then provided, with an emphasis on RDF- and AMR-to-Text generation. The chapter explored existing datasets and the languages they cover, common modeling approaches over time, and standard evaluation methods. In doing so, it highlighted the limitations of current approaches concerning low-resource languages.

The following chapters describe the research performed to address these limitations. Chapter 3 focuses on RDF-to-Text generation in low-resource Celtic languages using monolingual denoising and structured Soft Prompts. Chapter 4 deals with AMR-to-Text generation across both high- and low-resource languages through hierarchical fine-tuning and phylogeny-based language grouping. Chapter 5 proposes a referenceless multilingual evaluation framework based on Natural Language Inference.

# Chapter 3

# RDF-to-Text Generation of Celtic Languages

**Contents**



As previously discussed, most G2T research has focused exclusively on English, with a notable lack of parallel corpora in other languages being one of the main issues. This chapter addresses the first research question: Can RDF-to-Text generation be improved in low-resource languages with limited training examples by fine-tuning a model with phylogeny-inspired Soft Prompts?

## 3.1 Introduction

While subsection 2.2.3 presented the steady progress that has been made on the G2T generation task, it also highlighted how most of the progress has been made exclusively in English, with a few advances taking place in other high-resource languages like Russian. However, little research has been conducted on low-resource languages, which the data-intensive nature of the best-performing G2T approaches can partially explain.

Recent work in machine translation (Conneau et al., 2020; Lin et al., 2020) shows that fine-tuning large language models pre-trained on multiple languages helps compensate for data sparsity. Other studies have shown that lightweight fine-tuning techniques allow preserving language knowledge obtained from high-resource languages while transferring to low-resource languages. In particular, phylogenetic information (that is, information about how languages relate to each other based on their origins) has shown promising results in transfer learning for related languages in classification tasks like POS tagging, Named Entity Recognition, or Natural Language Inference (Faisal and Anastasopoulos, 2022). At the same time, Factorized Soft Prompts have demonstrated a good performance in transfer learning to low-resource languages in text generation tasks like summarization (Vu et al., 2022). Finally, some studies indicate that combining adaptation techniques with full fine-tuning might help mitigate the challenges of multilingual training (Pfeiffer et al., 2022).

This chapter focuses on G2T generation, where the input is a knowledge graph in RDF (Lassila and Swick, 1999) format and the output is a text verbalizing the graph in several languages from the Celtic family, including Irish (Gle), Welsh (Cym), and Breton (Bre). The method described here consists of two elements: a multilingual pre-trained language model, which is expected to provide general linguistic knowledge, and a task and phylogenetically informed factorized soft prompt, designed to learn language-specific weights. To train the model, two different strategies are employed: monolingual unsupervised denoising pre-training, which leverages the benefits of unlabeled data (essential for low-resource languages), and fine-tuning on a limited number of RDF-to-Text instances.

Leveraging the data made available by the WebNLG Challenge 2023 (Cripwell et al., 2023), the approach described in this chapter outperforms simple full fine-tuning and factorized soft prompts full fine-tuning without phylogeny information, both in terms of automatic metrics and human evaluation. Additionally, to analyze the impact of various components of the soft prompt, an ablation study was performed. Finally, an experiment was conducted to examine how the size of the RDF-to-Text fine-tuning data (from 0 to 1.5K) impacts generation performance.

## 3.2 Method

At the heart of this approach is a highly structured soft prompt that can be decomposed into multiple sub-prompts, each focused on a different aspect.

Inspired partly by the structure of the original T5 translation prompts (Raffel et al., 2020) (e.g., *Translate English to German*), the soft prompts are first divided into three main components: Task, Source, and Target. This selection of components is similar to a standard practice in Machine Translation architectures, such as mBART (Liu et al., 2020), M2M100 (Fan et al., 2021), and NLLB (NLLB Team et al., 2022), where both the Source and Target languages are specified to improve Zero-Shot performance.

The Source and Target components are further decomposed using phylogenetic language information. Each of them is split into Family, Genus, and Language sub-prompts to model phylogeny information. The resulting soft prompt is called *Phylogeny-Inspired Task-Source-Target* (PITST) soft prompt. The intuition is that using this prompt allows less-resourced languages to benefit from the training data of their related languages while mitigating the introduction of noise caused by the mixture of training data. Figure 3.1 shows the simplified phylogenetic tree used during training. The linearized RDF graphs are paired with the English language since the subjects, objects, and predicates of the RDF graphs are in English.

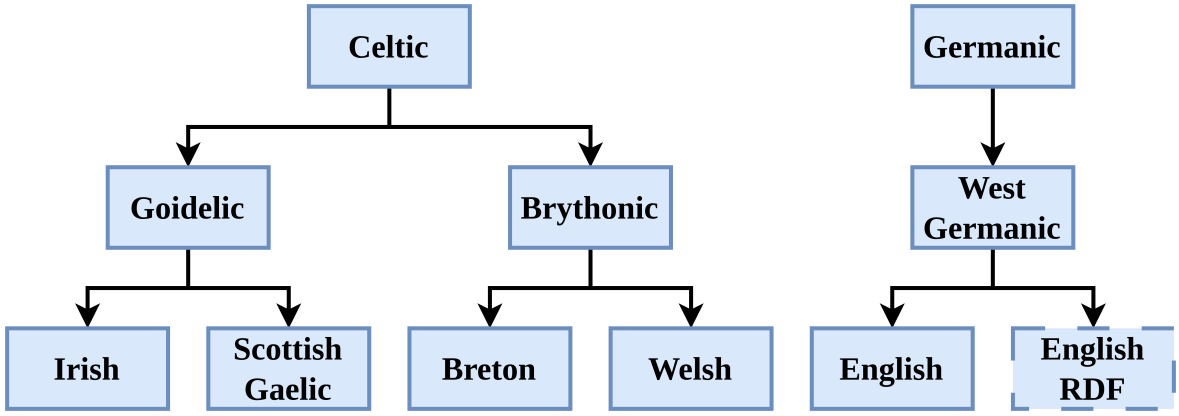



FIG. 3.1: Soft prompt phylogeny tree used during training.

To train and use this Soft Prompt, multiple steps are applied on a pre-trained multilingual model:

**Step 0: Language Model Adaptation.** Sometimes, the pre-training objective of a Model is not aligned with the natural text generation objective. For example, models based on T5 are generally pre-trained on the Span Corruption objective, which generates spans of text separated by sentinel tokens instead of plain natural text. This behavior is soon corrected when performing full model fine-tuning; however, lightweight approaches, such as Soft Prompts, can be more challenging to overcome. Lester et al. (2021) proved that further pre-training the base model on a language modeling task, like Prefix Language Modeling, benefits performance. Based on the better performance reported for the Masked Language Modeling (MLM) objective over causal language modeling by Raffel et al. (2020), particularly on the translation downstream tasks. In this research, the BERT-style MLM pre-training task from Raffel et al. (2020) is employed, rather than the Prefix Language Modeling (PLM) approach.

**Step 1: Unsupervised Pre-training of the Soft Prompt.** This step aims to train the language components of the soft prompt so that each of them captures as much language information relevant to their assigned language. Specifically, the whole soft prompt is trained on a mixture of unsupervised, monolingual tasks (Masked LM, Prefix LM, Suffix LM, Generation, and Deshuffling). The parameters used for each component are substituted based on the language of the training sample. Instances that belong to the same language family share the same Family sub-prompt but have different Genus and Language sub-prompts. Table 3.1 shows the possible values of each component, and Figure 3.2 shows an example input batch for this step.

| Component | Possible Values |
|---|---|
| Task | Masked LM, Prefix LM, Suffix LM, Deshuffling, Open Generation, Data-to-Text |
| Family | Germanic, Celtic |
| Genus | West Germanic, Goidelic, Britonic |
| Language | English, RDF, Irish, Scottish Gaelic, Breton, Welsh |

Tab. 3.1: Soft prompt possible values for each factorized component.

| | Task | Source | | | Target | | | Original Input Sequences | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Family | Genus | Lang. | Family | Genus | Lang. | | | | | | |
| Input Batch | Masked LM | Germanic | West Germanic | RDF | Germanic | West Germanic | RDF | <S> | Einstein | <P> | <mask> | <P> | Poland |
| | Prefix LM | Germanic | West Germanic | English | Germanic | West Germanic | English | Thank | you | for | <mask> | <pad> | <pad> |
| | Suffix LM | Celtic | Britonic | Welsh | Celtic | Britonic | Welsh | <mask> | honno | ? | <pad> | <pad> | <pad> |
| | Deshuffling | Celtc | Britonic | Breton | Celtic | Britonic | Breton | skuizh | ? | out | Ha | <pad> | <pad> |
| | Generate | Celtc | Goidelic | Irish | Celtic | Goidelic | Irish | Seo | <mask> | <pad> | <pad> | <pad> | <pad> |

Fig. 3.2: Soft prompt example batch for step 1 (Unsupervised Pre-training of the Soft Prompt).

**Step 2: Downstream Task Fine-tuning of the Soft Prompt.** Once the language components of the Soft Prompt have learned to perform the unsupervised tasks, the Task sub-prompt is trained on the downstream task (RDF-to-Text generation). Following Vu et al. (2022), one of the unsupervised task Soft Prompt components is used to initialize the new task Soft Prompt component. In this case, the Masked LM component is used since it is the closest one to the RDF-to-Text task. In this step, language components of the soft prompt continue being changed "based on the language of each training instance.

**Inference.** At inference time, the task and language sub-prompts are combined as required by the specific inference task (i.e., generating into Breton, Irish, or Welsh).

## 3.3 Data

For unsupervised training, Celtic and English monolingual data was extracted from multiple datasets available in the Huggingface Hub [33]. Specifically, data was collected from different OPUS corpora (Tiedemann, 2012) (Bible Corpus (Christodouloupoulos and Steedman, 2015), DGT, EUConst, GNOME, KDE4, OfisPublik (Tyers, 2009), OpenSubtitles (Lison and Tiedemann, 2016), Opus-100 (Zhang et al., 2020a), ParaCrawl, QED (Abdelali et al., 2014), Tatoeba, and Ubuntu), CC-100 (Conneau et al., 2020), CC-Aligned (El-Kishky et al., 2020), CC-Matrix (Schwenk et al., 2021), ECDC Steinberger et al. (2014), mC4, OSCAR (Suárez et al., 2019), TaPACo (Scherrer, 2020), TedTalks (Cettolo et al., 2012), UDHR, and Wikipedia.

To process the text, it is first split into sentences using SentenceSplitter [34] with the default English settings. Then, each sentence was normalized using TextaCy[35], by applying bullet point normalization, hyphenated words normalization, quotation marks normalization, Unicode

[33] https://huggingface.co/datasets
[34] https://github.com/mediacloud/sentence-splitter
[35] https://textacy.readthedocs.io/en/latest/

normalization, white space normalization, and HTML tag removal. Finally, the sentences were filtered using FastText Language Identification (Joulin et al., 2016b,a)[36], keeping only those above a 0.5 threshold. For the Celtic languages, as many samples as possible were collected, while for English, the number was limited to prevent it from overshadowing the other languages. Table 3.2 shows the number of samples available on each dataset used.

| Version | Train | Validation | Test |
|---|---|---|---|
| *Monolingual* | | | |
| Bre | 1 206 546 | 250 | 250 |
| Cym | 12 993 205 | 250 | 250 |
| Eng | 7 959 035 | 250 | 250 |
| Gle | 7 996 721 | 250 | 250 |
| Gla | 1 019 593 | 250 | 250 |
| *WebNLG* | | | |
| RDF-to-Bre | – | 1 399 | 2 280 |
| RDF-to-Cym | – | 1 665 | 1 779 |
| RDF-to-Eng | 35 426* | 4 464 | 5 150 |
| RDF-to-Gle | – | 1 665 | 1 779 |

TAB. 3.2: Soft prompt collected dataset for the experiments with number of instances per set. Although a large dataset of monolingual data was collected, only a small portion is used during training. *The English training WebNLG data was only used during the Zero-Shot ablation experiment.

## 3.4 Experiments

### 3.4.1 Training Process

The specific details of the training process are the following:

**Step 0: Language Model Adaptation.** $\text{mT5}_{\text{Large}}$ (Xue et al., 2021) was used as the base model[37] since it has been pre-trained in several languages, including English, Irish, Scottish Gaelic, and Welsh. Before training the Phylogeny-Inspired Soft Prompt, language model adaptation was performed for 30 000 steps on monolingual data for English, Breton, Irish, Scottish Gaelic, and Welsh as well as RDF triples from WebNLG. Once the LM Adaptation was completed, the Phylogeny-Inspired Soft Prompts were trained in two steps as follows.

**Step 1: Unsupervised Pre-training of the Soft Prompt.** This step takes 30 000 steps over the monolingual data for English, Breton, Irish, Scottish Gaelic, and Welsh, as well as the RDF triples from WebNLG.

**Step 2: Downstream Task Fine-tuning of the Soft Prompt.** The Task sub-prompt is further fine-tuned on the WebNLG task using the validation split of the English WebNLG dataset (Gardent et al., 2017) and human-written translations in Breton, Irish, and Welsh. This process takes five epochs or around 4 500 steps, keeping the best checkpoint every 500 steps.

[36] https://fasttext.cc/docs/en/language-identification.html
[37] https://huggingface.co/google/mt5-large

To account for the unbalanced distribution of samples in the datasets, the sampling strategy described in Devlin et al. (2019) was applied with $\alpha = 0.3$, which has been shown to perform best (NLLB Team et al., 2022). Table 3.3 accounts for that and other relevant hyperparameters used. The batch size was chosen to optimize the use of the GPUs. The learning rate was selected after a small exploratory experiment. The Soft Prompt size follows Vu et al. (2022), using around 50 tokens for the task and 50 for each language. Finally, the training steps follow Lester et al. (2021).

| **Hyperparameter** | **Value** |
|---|---|
| Base Model | mT5-Large |
| Vocabulary Size | ∼250K Tokens |
| Embedding Dimensions | 1 024 |
| Base Model Parameters | ∼1.22B |
| Total Prompt Parameters | ∼747K |
| Inference Prompt Parameters | ∼143 |
| Learning Rate | 0.0001 |
| Batch Size per GPU | 8 |
| Available GPUS | 2 Nvidia A40 |
| Sampling Temperature | 0.3 |
| ML Adaptation Steps | 30 000 |
| ML Adaptation Training Hours | ∼12 |
| Soft Prompt Pre-training Steps | 30 000 |
| Soft Prompt Pre-training Training Hours | ∼12 |
| Soft Prompt Fine-tuning Steps | ∼4 500 |
| Soft Prompt Fine-tuning Training Hours | ∼4 |

Tab. 3.3: Soft prompt hyperparameters.

### 3.4.2 Models

The proposed models are compared to a baseline obtained by applying simple full fine-tuning on mT5, one previous work with high performance in English, and two MT-based, upper-bound models.

**Simple Full Model Fine-tuning.** simple full fine-tuning on mT5 is performed. First, the language model adaptation is performed to attune the model to the target languages. It is then further fine-tuned for the downstream task.

**Control Prefixes.** The Control Prefixes model presented by Clive et al. (2022) is currently one of the best-performing strategies for the English WebNLG benchmark. This lightweight fine-tuning approach incorporates attribute-level parameters in various layers of T5, which indicate the semantic category of the input WebNLG RDF graph to enhance performance. For the baseline, a Control Prefixes variation is trained on the WebNLG validation data of all languages (Celtic and English).

**Machine Translation (MT).** Machine Translation is studied in two scenarios: a generate-and-translate scenario (NLG+MT), where the output of the best RDF-to-English generation system from the WebNLG Challenge 2020 (Guo et al., 2020a) is translated into the Celtic languages

using Machine Translation, and a translation-only scenario (Gold+MT), where the translation takes as input the references of the WebNLG dataset. These models serve as upper bounds, as they are trained on a much larger collection of parallel English-Celtic data, unlike the proposed models, which are trained on around 1.5K examples of aligned data per language. Furthermore, the Gold+MT model does not perform RDF-to-Text generation as it simply translates the English sentences of the WebNLG test set into Celtic. To perform the translations, a version of the system from Zhang et al. (2020a) trained only on Celtic and English data from the OPUS Corpora (Tiedemann, 2012) was used. It is worth noting that NLG+MT and the Gold+MT models require significantly more parallel data for training than the proposed method.

### 3.4.3 Ablation Experiments

Ablation experiments were performed to test the impact of the various sub-prompts (task, phylogeny data, source, and target language). The full PI-TST prompt is compared with five other prompts: the same prompt but without phylogeny information (TST), the same prompt without Source Language information (PI-TT), and three simplified prompts without phylogeny information, which are either unstructured (S) or model only two factors: Task and Target Language (TT) or Source and Target Language (ST). The size of the soft prompts is fixed at 140 tokens for all the experiments. When a task component was present, its size was fixed at 50 tokens, with the remaining components taking 90 tokens. All the language-related components on a soft prompt had their size distributed uniformly. Figure 3.3 shows the various prompts that were experimented with.

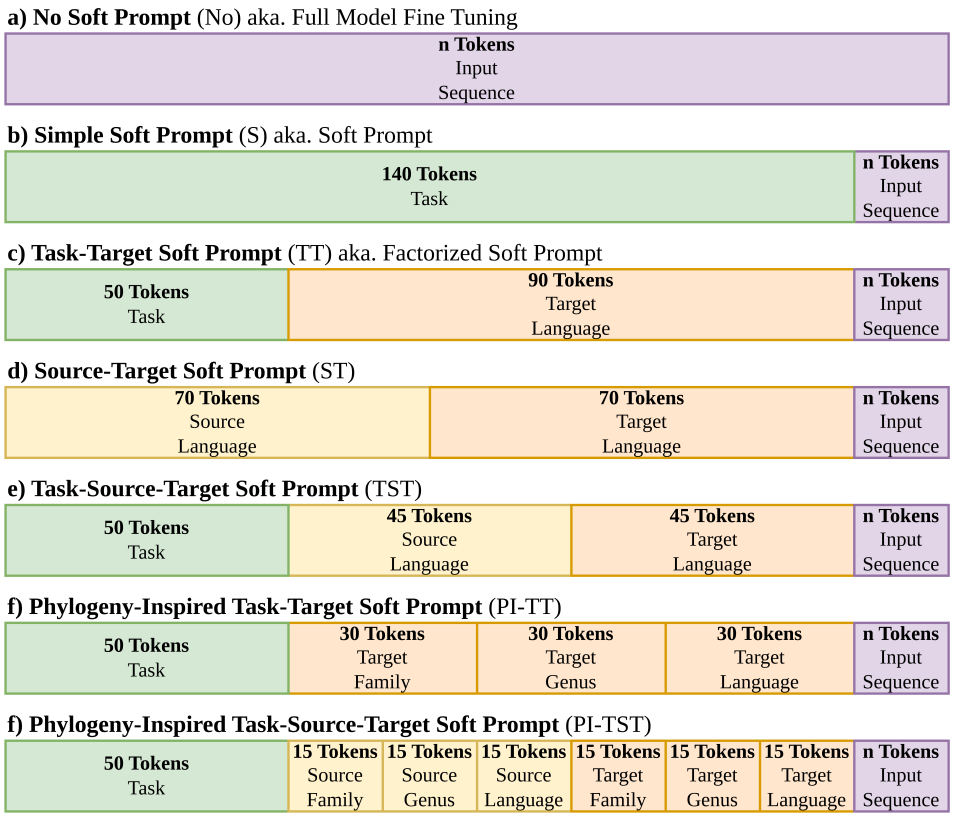



FIG. 3.3: Soft prompt variants and the composition of their input embeddings.

### 3.4.4 Training Data Experiments

**Zero-Shot.** The zero-shot capabilities of the final PI-TST model were tested by fine-tuning the task Soft Prompt only in English (either on the validation or training data) and testing it on Celtic languages.

**Training Samples.** This experiment tests the final PI-TST model by fine-tuning it using different numbers (100, 500, 1000) of randomly sampled elements from each language on the dataset.

## 3.5 Evaluation

### 3.5.1 Automatic Evaluation

**BLEU.** The corpus-level BLEU score (Papineni et al., 2002) was computed for each experiment using SacreBLEU (Post, 2018).

**Google BLEU.** Additionally, for each experiment, the sentence-level Google BLEU scores were computed(Wu et al., 2016b).

**LaBSE SBERT Cosine Similarity.** LaBSE (Feng et al., 2022) was used to obtain sentence embeddings for the generated text and the human reference. Then the sentence-level cosine similarity of both embeddings was computed as an automatic measurement of semantic accuracy. Given its implicit goal of being language-agnostic, this model was selected over others, which benefits experimentation on low-resource languages.

**Wilcoxon signed-rank test.** The Wilcoxon's signed-rank test (Wilcoxon, 1945) was used on the sentence-level metrics (Google BLEU and LaBSE Cosine Similarity) to evaluate whether the differences observed in different experiments are statistically significant. This approach was preferred over the paired Student's t-test since the results do not follow a normal distribution.

### 3.5.2 Human Evaluation

To perform human evaluation, 25 random input graphs from the test set were selected, ensuring a variety of sizes. The PI-TST model was then used to generate text from the same graphs in all the target languages. All 25 of those generated texts were provided to human evaluators, and they were asked to score them using readability, grammaticality, word order, and semantic adequacy. Each criterion was scored on a 1 to 3 Likert scale, where one is bad, two is medium, and three is good.

**Readability.** The evaluator was only given the model's output and asked if the generated text was understandable and reasonable in the language.

**Grammaticality.** The evaluator was only given the generated output of the model and asked if the morphology of the generated text was correct and if agreement constraints (e.g., verb/subject, noun/adjective) were respected.

**Word Order.** The evaluator was only given the model's generated output and asked if the word order of the generated text was correct and if a native speaker would come up with a text like that.

**Semantic Adequacy.** The evaluator was given the model's generated output and the human-written reference and asked if the generated text shared the same meaning as the human-written reference.

Colleagues who grew up in regions where the evaluated language is spoken were contacted to perform the human evaluation. Given the nature of the low-resource languages at hand, only a small number of evaluations were collected. Appendix A.1 provides more detail on the human evaluation process.

## 3.6 Results

### 3.6.1 Automatic Evaluation Results

Table 3.4 shows the results of the automatic evaluation.

| | BLEU Score ↑ | | | | Google BLEU Score ↑* | | | | LaBSE Cosine Similarity ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Experiment** | **Bre** | **Cym** | **Eng** | **Gle** | **Bre** | **Cym** | **Eng** | **Gle** | **Bre** | **Cym** | **Eng** | **Gle** |
| *Machine Translation* | | | | | | | | | | | | |
| NLG+MT | *13.08* | *20.24* | *53.98* | *18.09* | *17.74* | *27.49* | *49.64* | *24.86* | *72.96* | *89.90* | *95.05* | *87.76* |
| Gold+MT | *19.81* | *49.04* | *100.00* | *32.09* | *23.04* | *51.82* | *100.00* | *36.44* | *76.23* | *94.80* | *100.00* | *92.56* |
| *Baselines* | | | | | | | | | | | | |
| Control Prefixes | 12.23 | 13.33 | **51.61** | 8.17 | 16.37 | 18.76 | 47.77 | 13.59 | 80.52 | 79.41 | **94.52** | 73.12 |
| Full Fine-tuning | 16.49 | 18.83 | 46.40 | 14.16 | 21.36 | 24.36 | 43.62 | 20.09 | 82.56 | 86.02 | 92.35 | 82.49 |
| *Final* | | | | | | | | | | | | |
| PI-TST | **18.15** | **20.60** | 49.15 | **15.64** | **22.57** | **25.95** | 46.09 | **21.23** | **84.09** | **87.72** | 93.65 | **84.68** |

Tab. 3.4: Soft prompts automatic evaluation results. For Google BLEU and cosine similarity, the results without a statistically significant difference from the final PI-TST model ($p > 0.05$) are underlined. The English values on the machine translation rows are the scores obtained by the RDF-to-EN model and the gold references, i.e., in this case, translation is not used. *Since the sentence-level Google BLEU score is used for statistical significance analysis, here the average of the sentence-level score is presented, instead of the corpus-level one.

**PI-TST Outperforms the Baselines.** The PI-TST proposal outperforms simple mT5 full fine-tuning and the state-of-the-art Control Prefixes models fine-tuned on Celtic data. For Breton and Welsh, PI-TST even outperforms the BLEU score of the NLG+MT approach, with the advantage that this model does not require the amount of parallel translation data that training the MT model requires, which is not always available for low-resource languages. Furthermore, the NLG model of the NLG+MT baseline was trained on all 32K samples of the full English WebNLG, while PI-TST is only trained on validation data, which is significantly smaller. It is worth noting that, for Breton, which is the most under-resourced of the Celtic languages evaluated, the proposed method even comes close to the Gold+MT BLEU score and surpasses its LaBSE Cosine Similarity score. As the data used to pre-train mT5 does not include any Breton, this suggests that the fine-tuning approach produces larger improvements on languages that were not seen during the base model's pre-training.

**The Effect of Source Information.** The ablation results in Table 3.5 show that the two best-performing models (PI-TST, ST) include a source and target sub-prompt, which suggests that, similar to the control tokens used in multilingual machine translation, the source and target sub-prompts help structure the representation space and guide learning. The conjecture is that having both Source and Target sub-prompts (rather than just Target) helps the model differentiate between the unsupervised monolingual step (Step 1), where Source and Target prompts refer to the same language and the second fine-tuning step where the Source and Target prompt refers to different languages (Source: RDF, Target: Celtic). On the other hand, the results showed that the TST model without PI has much lower performance than ST, likely due to a trade-off between prompts and prompt size: 70 tokens for the Source token in ST vs. 45 in TST.

**The Effect of Phylogenetic Information.** Like PI-TST, the Phylogeny-Inspired Task-Target (PI-TT) model outperforms simple full fine-tuning in all languages, confirming the positive impact of phylogeny information.

**Languages not Seen During Pre-Training of the Original Encoder-Decoder (mT5).** For Breton, the only language not seen during the pre-training of mT5, the PI-TT model outperforms TT, indicating that phylogeny information benefits under-resourced languages.

**Source and Phylogeny Prompts.** Comparing models across these two dimensions shows that, while adding either a phylogeny or a source sub-prompt does not always improve performance (both TST and PI-TT underperform TT), adding both does help (PI-TST outperforms all other models).

**Zero-Shot.** Table 3.5 shows that using the PI-TST model in a zero-shot setting reaches equivalent results on Celtic languages as a simple Soft Prompt model trained on all Celtic languages.

| | **BLEU Score ↑** | | | | **Google BLEU Score ↑*** | | | | **LaBSE Cosine Similarity ↑** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Experiment** | **Bre** | **Cym** | **Eng** | **Gle** | **Bre** | **Cym** | **Eng** | **Gle** | **Bre** | **Cym** | **Eng** | **Gle** |
| *Soft Prompt* | | | | | | | | | | | | |
| S | 9.63 | 1 1.01 | 48.48 | 10.36 | 13.41 | 15.18 | 44.73 | 14.18 | 79.84 | 86.42 | 93.51 | 82.49 |
| TT | 17.70 | 19.94 | 48.30 | 15.58 | 21.95 | 25.32 | 45.26 | 21.04 | 83.21 | 87.59 | 93.60 | 84.66 |
| ST | 17.89 | 19.94 | 49.18 | 15.58 | 22.24 | 25.34 | 45.73 | 20.88 | 83.72 | 87.53 | 93.55 | 84.47 |
| TST | 16.28 | 18.49 | 47.29 | 15.39 | 21.33 | 24.19 | 44.82 | 20.94 | 82.21 | 86.46 | 93.04 | 84.16 |
| PI-TT | 17.43 | 19.41 | 48.32 | 15.23 | 22.16 | 25.28 | 45.29 | **21.48** | 83.55 | 87.34 | 92.90 | 84.35 |
| *Training Samples* | | | | | | | | | | | | |
| 100 Samples | 12.42 | 13.61 | 38.42 | 10.66 | 17.12 | 18.98 | 38.09 | 15.66 | 77.58 | 81.15 | 89.74 | 78.56 |
| 500 Samples | 14.31 | 14.95 | 43.70 | 12.60 | 19.08 | 20.68 | 42.12 | 16.99 | 79.92 | 82.54 | 91.30 | 79.84 |
| 1000 Samples | 15.34 | 18.29 | 47.18 | 13.91 | 20.29 | 24.02 | 44.29 | 19.32 | 81.99 | 86.44 | 92.47 | 82.53 |
| *Zero-Shot* | | | | | | | | | | | | |
| English Validation | 9.81 | 11.85 | 48.36 | 9.69 | 13.79 | 16.88 | 45.04 | 13.96 | 78.57 | 83.29 | 93.26 | 82.19 |
| English Training | 9.57 | 11.27 | 48.09 | 10.36 | 13.49 | 16.19 | 44.95 | 14.58 | 79.19 | 83.70 | 92.94 | 81.04 |
| *Final* | | | | | | | | | | | | |
| PI-TST | **18.15** | **20.60** | **49.15** | **15.64** | **22.57** | **25.95** | **46.09** | 21.23 | **84.09** | **87.72** | **93.65** | **84.68** |

TAB. 3.5: Soft prompt ablations automatic evaluation results. For Google BLEU and cosine similarity, the results without a statistically significant difference from the final PI-TST model ($p > 0.05$) are underlined. *Since the sentence-level Google BLEU score is used for statistical significance analysis, here the average of the sentence-level score is presented, instead of the corpus-level one.

**Size of the Training Data.** Figure 3.4 shows the performance of the PI-TST models when fine-tuned with varying amounts of graph-text data. With only 1 000 samples per language, PI-TST outperforms Full Model fine-tuning in English and performs on par with the Celtic languages.

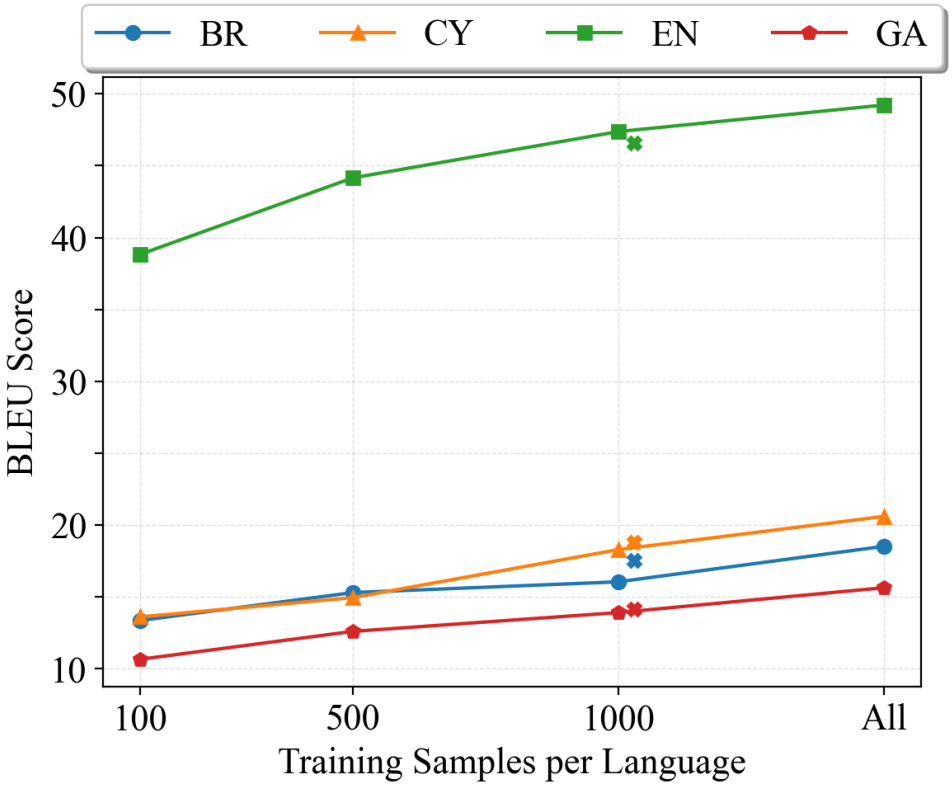



FIG. 3.4: Soft prompt BLEU comparison by number of training samples per language. The × mark indicates the score of Full Model Fine-tuning.

**Statistical Significance.** Table 3.6 presents the statistical significance between each experiment and the final proposal, PI-TST. While some of the ablation experiments produce results that are not statistically different from the proposed method, an argument in its favor can be made against those other approaches, as PI-TST provides much more controllability and flexibility given its complex soft prompt. The extreme modularity of this proposal gives it an edge over the ablation studies. It is also notable that, where the average Google BLEU score of ablation experiments outperformed the proposal (Irish PI-TT), the difference was not statistically significant. Finally, the difference in the Google BLEU score between the proposal and the Breton Gold+MT is not statistically significant, despite the former (and more data-intensive) approach having a higher average.

| Experiment | Google BLEU Score ↑* | | | | LaBSE Cosine Similarity ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| | **Bre** | **Cym** | **Eng** | **Gle** | **Bre** | **Cym** | **Eng** | **Gle** |
| *Machine Translation* | | | | | | | | |
| NLG+MT | 0.0000 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Gold+MT | 0.2700 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| *Baselines* | | | | | | | | |
| Control Prefixes | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Full Fine-tuning | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| *Soft Prompt* | | | | | | | | |
| S | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0121 | 0.0000 |
| TT | 0.0007 | 0.0089 | 0.0135 | 0.1101 | 0.0060 | 0.4284 | 0.5420 | 0.4962 |
| ST | 0.0249 | 0.0048 | 0.1443 | 0.0124 | 0.0626 | 0.1318 | 0.4616 | 0.0467 |
| TST | 0.0000 | 0.0000 | 0.0000 | 0.0089 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| PI-TT | 0.0166 | 0.0011 | 0.0020 | 0.1358 | 0.0013 | 0.0003 | 0.0000 | 0.0313 |
| *Training Samples* | | | | | | | | |
| 100 Samples | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 500 Samples | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1000 Samples | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| *Zero-Shot* | | | | | | | | |
| English Validation | 0.0000 | 0.0000 | 0.0004 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| English Training | 0.0000 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Tab. 3.6: Soft prompt Wilcoxon signed-rank test p-values. For Google BLEU and cosine similarity, the results without a statistically significant difference from the final PI-TST model ($p > 0.05$) are underlined. *Since the sentence-level Google BLEU score is used for statistical significance analysis, here the Average of the sentence-level score is presented, instead of the corpus-level one.

### 3.6.2 Human Evaluation Results

When asked where they learned the language, 4 of the evaluators answered *Home*, 2 answered *School and Home*, and 3 answered *School*. When asked how they considered their proficiency in the language, 8 of the evaluators answered *Good* and one answered *Medium*. Table 3.7 shows the results of their evaluation of the PI-TST model. This evaluation indicates that the model produces acceptable text in terms of readability, grammar, and word order for all Celtic languages. It also shows that, for English and Irish, the quality of the semantic adequacy is past the middle point.

| **Criteria** | **Bre** | **Cym** | **Eng** | **Gle** |
|---|---|---|---|---|
| Annotators | 3 | 2 | 2 | 2 |
| Readability ↑ | 0.84 | 0.59 | 0.98 | 0.58 |
| Grammaticality ↑ | 0.84 | 0.73 | 0.92 | 0.71 |
| Word Order ↑ | 0.84 | 0.79 | 0.97 | 0.65 |
| Semantic Adequacy ↑ | 0.42 | 0.32 | 0.77 | 0.53 |

Tab. 3.7: Soft prompt human evaluation results converted to a 0-to-1 scale from the original 3-point Likert scale.

Tables 3.8 and 3.9 display generated text in a low-resource language (Welsh) and a high-resource language (English). The examples are taken from the 1st (best), 3rd (medium), and 5th (worst) quintiles, based on the average of the four human evaluation metrics. Appendix A.2 provides similar examples for the other low-resource Celtic languages (Breton and Irish).

| Sample | R | G | WO | SA | Q |
|---|---|---|---|---|---|
| **Graph:**<br>Nie Haisheng \| birthPlace\| Zaoyang<br>Nie Haisheng \| birthDate \| 1964-10-13<br>**Generation (Cym):**<br>Ganed Ni Haisheng ar 13 Medi 1964 yn Zaoyang.<br>**Reference (Cym):**<br>Ganwyd Nie Haisheng ar Hydref 13, 1984 yn Zaoyang.<br>**Generation (Eng MT*):**<br>Ni Haisheng was born on 13 September 1964 in Zaoyang.<br>**Reference (Eng):**<br>Ganed Ni Haisheng ar 13 Medi 1964 yn Zaoyang. | 1.00 | 1.00 | 1.00 | 0.75 | 1st |
| **Graph:**<br>Harold French \| birthPlace \| UK<br>**Generation (Cym):**<br>Ganed Harold French yn yr Unol Daleithiau.<br>**Reference (Cym):**<br>Cafodd Harold French ei eni yn y DU.<br>**Generation (Eng MT*):**<br>Harold French was born in the United States.<br>**Reference (Eng):**<br>Harold French is born in the UK. | 0.75 | 1.00 | 1.00 | 0.25 | 3rd |
| **Graph:**<br>The Hobbit \| literaryGenre \| High fantasy<br>**Generation (Cym):**<br>Mae The Hobbit yn y genre cerddorol sydd i'w gael yn y cyfres.<br>**Reference (Cym):**<br>Genre yr Hobbit yw ffantasi uchel.<br>**Generation (Eng MT*):**<br>The Hobbit is in the musical genre found in the series.<br>**Reference (Eng):**<br>The Hobbit literary genre is high fantasy. | 0.00 | 0.50 | 0.25 | 0.00 | 5th |

Tab. 3.8: PI-TST Welsh generation examples from the human evaluation with the average score across evaluators. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) (Q) based on the average of the four human evaluation metrics: readability (R), grammaticality (G), word order(WO), and semantic adequacy (SA). *The MT model may have altered differences in the original Welsh generation.

| Sample | R | G | WO | SA | Q |
|---|---|---|---|---|---|
| **Graph:**<br>Harold French \| birthPlace \| UK<br>**Generation (Eng):**<br>Harold French was born in the UK.<br>**Reference (Eng):**<br>Harold French is born in the UK. | 1.00 | 1.00 | 1.00 | 1.00 | 1st |
| **Graph:**<br>Terence Rattigan \| deathYear \| 1977-01-01<br>**Generation (Eng):**<br>Terence Rattigan died on January 1st, 1977.<br>**Reference (Eng):**<br>Terence Rattigan died in 1977. | 1.00 | 1.00 | 1.00 | 0.50 | 3rd |
| **Graph:**<br>Nie Haisheng \| birthPlace\| Zaoyang<br>Nie Haisheng \| birthDate \| 1964-10-13<br>**Generation (Eng):**<br>Ni Haisheng was born on 10th October 1964 in Zaoyang.<br>**Reference (Eng):**<br>Nie Haisheng was born on October 13, 1984 in Zaoyang. | 1.00 | 0.75 | 1.00 | 0.25 | 5th |

Tab. 3.9: PI-TST English generation examples from the human evaluation with the average score across evaluators. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the four human evaluation metrics: Readability (R), grammaticality (G), word order(WO), and semantic adequacy (SA).

## 3.7 Conclusion

In this chapter, the first research question was addressed by introducing a fine-tuning strategy for RDF-to-Text generation in low-resource Celtic languages that leverages phylogeny-informed Soft Prompts. By integrating information about the task and details about the language family, genus, and specific language of both the source and target languages in structured prompt components and combining them with monolingual unsupervised pre-training, this approach demonstrated that the Phylogeny-Inspired Task-Source-Target (PI-TST) soft prompt approach improves the quality of generation.

The results showed that this method outperforms both a simple complete fine-tuning baseline and the state-of-the-art Control Prefixes method, even when using a smaller number of RDF-to-Text training examples. Notably, the model performs well in Breton, a language not seen during the base model's pre-training, underlining the benefits of incorporating phylogenetic structure for transfer learning in unseen or extremely low-resource settings.

These findings validate the initial hypothesis: RDF-to-Text generation in low-resource languages can be meaningfully improved with limited labeled data by fine-tuning with structured Soft Prompts. These findings open up promising directions for scalable and language-aware natural language generation (NLG) techniques, especially in underrepresented linguistic contexts.

# Chapter 4

# AMR-to-Text Generation of High- and Low-resource Languages

**Contents**



The promising results obtained using phylogenetic information motivated us to further explore its application to similar problems. This chapter addresses the second research question: Can text generation from Abstract Meaning Representation (AMR) graphs be improved using phylogenetic information to guide a model's training process in high- and low-resource languages?

## 4.1 Introduction

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a representation language used to encode the meaning of sentences. AMR-to-Text generation is the task of verbalizing the meaning encoded by an AMR graph. While there has been constant progress on this task for the English language (Hoyle et al., 2021; Ribeiro et al., 2021b,c; Bevilacqua et al., 2021) and some other high-resource (HR) and medium-resource (MR) languages (Fan and Gardent, 2020; Ribeiro et al., 2021a; Xu et al., 2021; Martínez Lorenzo et al., 2022; Sobrevilla Cabezudo and Pardo, 2022), not much attention has been given to this task in low-resource (LR) languages.

Previous work on machine translation (MT) exposes a complex trade-off between high- and low-resource languages during training. While Koehn and Knowles (2017) showed that neural

MT models have a steep learning curve leading to poor performance in low-resource scenarios, Lin et al. (2020) and Aharoni et al. (2019) demonstrate that multilingual training mitigates this effect. Conversely, Conneau et al. (2020) observe that the noise resulting from multilingual training negatively affects HR languages, while NLLB Team et al. (2022) show that curriculum learning (Bengio et al., 2009) can help reduce over-fitting on LR languages.

Phylogenetic knowledge has sometimes been used to handle this trade-off, both in multilingual NLU tasks such as dependency parsing, part of speech tagging, and natural language inference (Faisal and Anastasopoulos, 2022) and in NLG tasks such as G2T generation (see Chapter 3). Recent work (Meng and Monz, 2024) has also shown that training on closely related languages facilitates transfer, while training on distant languages has a regularization effect. Finally, Parameter-Efficient Fine-Tuning approaches have proven helpful in learning new tasks and languages for text generation of LR languages (Vu et al., 2022) while keeping memory requirements low during training.

This chapter focuses on AMR-to-Text generation and proposes two techniques to improve transfer from high- to low-resource languages while preserving performance in HR languages. First, iteratively refining a multilingual model into a set of monolingual models using Low-Rank Adapters (LoRA) (Hu et al., 2022). With the hypothesis that this promotes cross-lingual transfer, limits the impact of data sparsity for LR languages, and reduces over-fitting of HR languages as the monolingual models are trained last. Second, this training curriculum relies on a tree structure whose nodes indicate which languages are included in the training data at each iteration step. Using phylogenetic knowledge, high- and low-resource languages are grouped, either with closely related or distant languages. In this way, the chapter investigates how using different phylogenetic-based training strategies impacts performance.

## 4.2 Method

To mitigate the effects of data scarcity (over-fitting) and multilingual training (noise), this chapter proposes a variation of curriculum learning that leverages both phylogenetic knowledge and the modularity and memory efficiency of LoRAs to iteratively refine a base multilingual model into a set of monolingual models.

**Base Model.** The base model is $\text{mT5}_{\text{large}}$ (Xue et al., 2021)[38], a multilingual encoder-decoder model extended with LoRAs to support modular Parameter-Efficient Fine-Tuning and 4-bit quantization to reduce memory footprint during training.

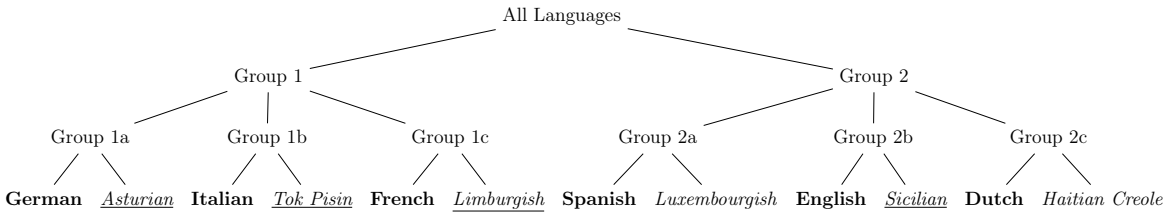
**Refining Models.** The goal is to learn 12 monolingual models by iteratively fine-tuning a model in four steps. In the first step (Level 0), the base model ($\text{mT5}_{\text{large}}$) is fine-tuned on all 12 languages using a single LoRA fine-tuning. The resulting model, which is created by merging ($\text{mT5}_{\text{large}}$)'s weights with the LoRA, is then fine-tuned on two sets of 6 languages, yielding two 6-language models, each trained with a separate LoRA module (Level 1). This process is repeated twice: first, fine-tuning the two 6-language models into six bilingual models (Level 2) and then fine-tuning each bilingual model into 12 monolingual models (Level 3). Algorithm 1 provides more detail on this process.

[38] https://huggingface.co/google/mt5-large

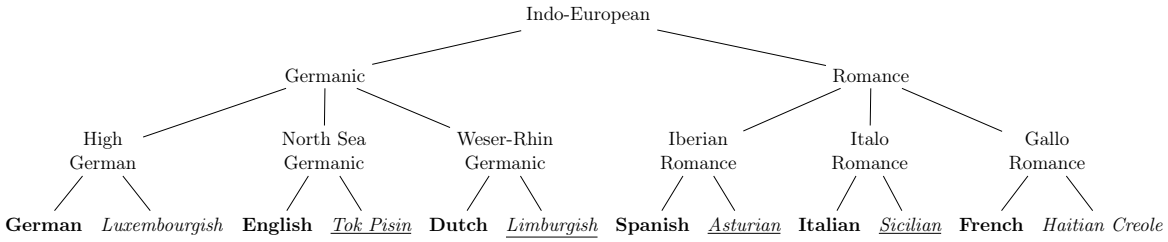
**Algorithm 1** HQL training algorithm simplified.

```
Load 4-bit Quantized Base Model
Load Training Hierarchy
for Level in Training Hierarchy do
    for Group in Level do
        Load data of all languages in Group
        if Level not 0 then
            Load relevant LoRAs from previous Levels
            Merge LoRAs with model
        end if
        Add new LoRA
        Train LoRA on relevant data
        Save current Group LoRA
    end for
end for
```

**Choosing Language Groups.** The proposed training strategy is structured as a four-level tree, where each node determines the set of languages used for fine-tuning the parent model. The specific languages studied are German (Deu), Luxembourgish (Ltz), English (Eng), Tok Pisin (Tpi), Dutch (Nld), Limburghish (Lim), Spanish (Spa), Asturian (Ast), Italian (Ita), Sicilian (Scn), French (Fra), and Haitian Creole (Hat). Based on previous works, the effect of the two training hierarchies shown in Figure 4.1 is explored.




(a) Distant Languages Hierarchy (DLH)




(b) Phylogenetic Tree Hierarchy (PTH)

Fig. 4.1: HQL training hierarchies tested. The top one (DLH) maximizes the language difference within nodes of each level. The bottom one (PTL) minimizes the language difference within nodes of each level. High-resource languages are in **bold**, low-resource languages are in *italics*, and languages unseen by the pre-trained base model are underlined.

Meng and Monz (2024) showed that balanced data from distant languages during training can be a regularizing factor. Accordingly, the first strategy is to increase the average distance between languages for each node in the training hierarchy. This strategy produces the Distant Languages Hierarchy (DLH) depicted in 4.1a.

Conversely, multiple previous studies have pointed to the benefits of training multilingual models on closely related languages. Based on this, the second training hierarchy follows the phylogenetic tree shown in 4.1b, where at each level of the hierarchy, the corresponding LoRA module is trained on smaller, less diverse, and more closely related groups of languages. Under this Phylogenetic Tree Hierarchy (PTHQL) approach, the expectation is to increase the transfer learning and reduce the noise of other languages as training progresses.

## 4.3 Data

### 4.3.1 Training Data

The 2020 AMR release 3.0 dataset (Knight et al., 2020)[39], also known as LDC2020T02, includes 55.6K gold-quality AMR-Text pairs, where a human created both the graph and the English text. A silver-quality training dataset was created for all target languages using machine translation and language identification filtering on the English gold-quality data. First, the English texts are translated to a target language using a 4-bit quantized NLLB-3.3B model (NLLB Team et al., 2022)[40]. Second, the machine-translated texts are filtered using the GlotLID (Kargaran et al., 2023)[41] language identification model by removing all instances with a score below 0.5. Third, the top 31K instances are retained for each language so that the quantity of training data is the same for all languages. This method produces a dataset of 31K (gold AMR, machine-translated texts) for each of the target languages, except English, where texts are human-written.

Additionally, a small parallel dataset is created for all target languages where the AMRs are silver and the texts are human-written. The text for this dataset is derived from the FLORES-200 dataset of parallel texts (NLLB Team et al., 2022). The silver AMR graphs are obtained by parsing the English texts of the dataset using AMR3-structbart-L (Drozdov et al., 2022)[42]. Since FLORES-200 does not include training data, the validation split was used for training, and the test split was divided in half to create two smaller validation and test sets.

### 4.3.2 Test Data

English, German, Spanish, and Italian were evaluated on gold-quality AMR-Text pairs on LDC2020T07 (Damonte and Cohen, 2018, 2020)[43], a subset of the 2020 AMR release 3.0 with gold AMR graphs and human-translated and corrected texts. For the remaining eight languages, the sub-test set from FLORES-200 (silver AMR, human-written text) pairs was used. The use of silver AMR graphs paired with human-verified sentences was preferred, despite the possibility of generating (gold AMR, machine-translated texts) pairs from the 2020 AMR release 3.0.

[39] https://catalog.ldc.upenn.edu/LDC2020T02
[40] https://huggingface.co/facebook/nllb-200-3.3B
[41] https://github.com/cisnlp/GlotLID
[42] https://github.com/IBM/transition-amr-parser/
[43] https://catalog.ldc.upenn.edu/LDC2020T07

The rationale behind this decision was that the noise introduced by the AMR parser when producing the silver AMR graphs would be more consistent across languages, making evaluation fairer. In contrast, the noise of machine-translated silver sentences would vary across languages, given the uneven performance of machine translation models. Furthermore, by comparing it against real human-written text, the quality of the generated text can be more accurately assessed. Table 4.1 summarizes the size and type of data.

| | Quality | | Instances per Language | | |
|---|---|---|---|---|---|
| **Dataset** | **AMR** | **Text** | **Train** | **Test** | **Valid** |
| FLORES-200 | Silver | Gold | 997 | 506 | 506 |
| English AMR 3.0 | Gold | Gold | N/A | 1 371 | N/A |
| Translated AMR 3.0 | Gold | Silver | 30 000 | 1 000 | 1 000 |

Tab. 4.1: HQL preprocessed datasets.

## 4.4 Experiments

### 4.4.1 Training Process.

All the experiments use $\text{mT5}_{\text{large}}$ as the underlying base model via the transformers [44] library. The PEFT [45] library was used to handle the QLoRA implementation. The model was quantized to 4-bit precision for memory efficiency.

Following (Dettmers et al., 2023), the QLoRA was applied to all linear layers of the model, which improves performance. Since Hu et al. (2022) suggests that new languages and tasks might require much higher ranks, since the proposed models need to learn an entirely new task (AMR-to-Text vs Spam Correction) and generate in scarcely seen and previously unseen languages, the Rank and Alpha were set to 256 using Rank-Stabilized scaling. The base model contained around 1.2B parameters, and introducing LoRA adds almost 300M new trainable parameters.

A batch size of 8 was used, selecting a power of 2, which benefits the training speed. A maximum length per training instance of 256 tokens was selected, similar to the values chosen by Ribeiro et al. (2021a), which implied the truncation of around 8% of tokens on the input sequence but does not affect the output sequences.

Each model was trained on the same amount of data to factor out the impact of training data size. Starting with 30 997 distinct instances for each language and training for one epoch on each level of the training hierarchy. Thus, L0 models are trained on 371 964 (30 997 $\times$ 12) unique instances, L1 models on 185 982 (30 997 $\times$ 6) instances, L2 on 61 994 (30 997 $\times$ 2) instances, and L3 on 30 997 instances. Hence, by the end of the training, each monolingual model has seen 650 937 (30 997 $\times$ 21) instances, with unique instances seen up to 4 times across models, equivalent to training four epochs on the full dataset.

It is worth noting that, given the modularity of QLoRAs and how the intermediate levels can be reused to train the new ones, the total number of instances used for training all 12 monolingual models is 1 487 856. In comparison, full fine-tuning 12 monolingual models that have seen 650 937 instances would require training on 7 811 244 (650 937 $\times$ 12) instances.

[44] https://huggingface.co/docs/transformers
[45] https://huggingface.co/docs/peft

This approach is used to train the Distant Languages Hierarchy (DLH) and the Phylogenetic Tree Hierarchy (PTH). Table 4.2 includes more details about the hyperparameters.

| **Hyperparameter** | **Value** |
|---|---|
| Dataset | AMR3.0 + Flores200 |
| Max sequence length | 256 tokens |
| Batch Size | 8 |
| Unique Instances per Language | 30 997 |
| Total Unique Instances | 371 964 |
| Epochs per Level (L) | 1 |
| Instances (L0) | 371 964 |
| Instances (L1) | 185 982 |
| Instances (L2) | 61 994 |
| Instances (L3) | 30 997 |
| Total Seen Instances | 650 937 |
| Real Total Instances | 1 487 856 |
| Checkpoints | Every 500 batches |
| Optimizer | Adafactor |
| Scheduler | Linear |
| Learning Rate | 5e-5 |
| Base Model | google/mt5-large |
| Base Model Parameters | 1.2B |
| LoRA Parameters | 293M |
| LoRA Rank | 256 |
| LoRA Alpha | 256 |
| LoRA Dropout | 0.05 |
| LoRA Scailing | Rank-Stabilized |
| LoRA Targets | All linear layer |
| Quantization | BnB 4-bit |

Tab. 4.2: HQL hyperparameters.

### 4.4.2 Models

#### Previous Works

*F&G* (Fan and Gardent, 2020) is an encoder-decoder multilingual model that supports 21 high- and medium-resource languages. The encoder includes structural embeddings, and the model was fine-tuned on (silver AMR, gold text) pairs with data sizes ranging from 400K to 8.2M pairs, depending on the target language.

*Ribeiro* (Ribeiro et al., 2021a) traine a $\text{mT5}_{\text{base}}$ model that supports 4 HR languages. It was fine-tuned on millions of (silver AMR, gold text) and tens of thousands of (gold AMR, silver text) pairs for each target language.

*Xu* (Xu et al., 2021) separately trained three transformers on three HR languages. First, they pre-trained on six tasks (AMR-to-Eng, Eng-to-AMR, Eng-to-X, X-to-English, AMR-to-X, and X-to-AMR) on millions of (silver AMR, gold text) pairs. Then they fine-tuned on two tasks (AMR-to-X and Eng-to-X) using 36.5K (gold AMR, gold Eng text / silver X text).

*Martinez* (Martínez Lorenzo et al., 2022) trained an $\text{mBART}_{\text{large}}$ model separately on four HR languages. In particular, the version trained on plain AMR inputs was used. This model was trained for up to 30 epochs on 55K (gold AMR, silver text) pairs.

#### Baselines

*Monolingual QLoRA (MonoQL).* Twelve monolingual models were obtained by fine-tuning $\text{mT5}_{\text{large}}$ on each language separately using LoRA. The expectation is that this model performs worse than the proposed HQL model, particularly on LR languages, due to the limited training data, which can lead to either a lack of generalization or to overfitting. Each final model of the HQL approach has seen 650 937 instances during training. To allow for a fair comparison, each *MonoQL* model is trained with that many instances.

*Multilingual QLoRA (MultiQL).* Fine-tuned $\text{mT5}_{\text{large}}$ using LoRA on data from all 12 languages. The expectation is that this model will perform worse than the proposed HQL models, due to the noise introduced by the language mix. Since training all the HQL models requires a total of 1 487 856 instances, this multilingual model trains up to that many instances.

*Generate and Translate (Gen&Trans).* AMR-to-English is performed using the English *MonoQL*. Then, it is translated into the target languages using the same model used to generate the silver data (4-bit quantized NLLB-3.3B). The expectation is that this model mirrors the uneven quality of machine translation models, performing well in high-resource (HR) languages but less so in low-resource (LR) languages.

#### *Proposed HQLs*

*Distant Languages Hierarchical QLoRA (DLHQL).* Multiple LoRAs trained using the proposed iterative curriculum learning with the distant language hierarchy from 4.1a. The expectation is that this model showcases the regularizing effect of distant languages.

*Phylogenetic Tree Hierarchical QLoRA (PTHQL).* Multiple LoRAs trained using the proposed iterative curriculum learning with the phylogenetic tree from 4.1b to guide the training hierarchy. The expectation is that this model increases transfer learning and reduces the noise of other languages as training progresses.

## 4.5 Evaluation

Following NLLB Team et al. (2022), evaluation is performed with BLEU. This simple surface-based metric does not rely on training data, which is an advantage when dealing with multiple languages, particularly those with low resources. The scores with SacreBLEU (Post, 2018)[46] and the default settings (including *13a* tokenizer) for comparability with previous works. Chrf++ and BLEURT [47] are also reported. However, the discussion of results focuses mainly on BLEU, given its widespread use in the past and it being the only metric available on all previous works that use the same test sets.

Statistical testing is computed via paired bootstrap resampling (Koehn, 2004) for BLEU and ChrF++ and the Wilcoxon signed-rank test (Wilcoxon, 1945) for BLEURT-20.

[46] https://github.com/mjpost/sacrebleu
[47] https://github.com/google-research/bleurt

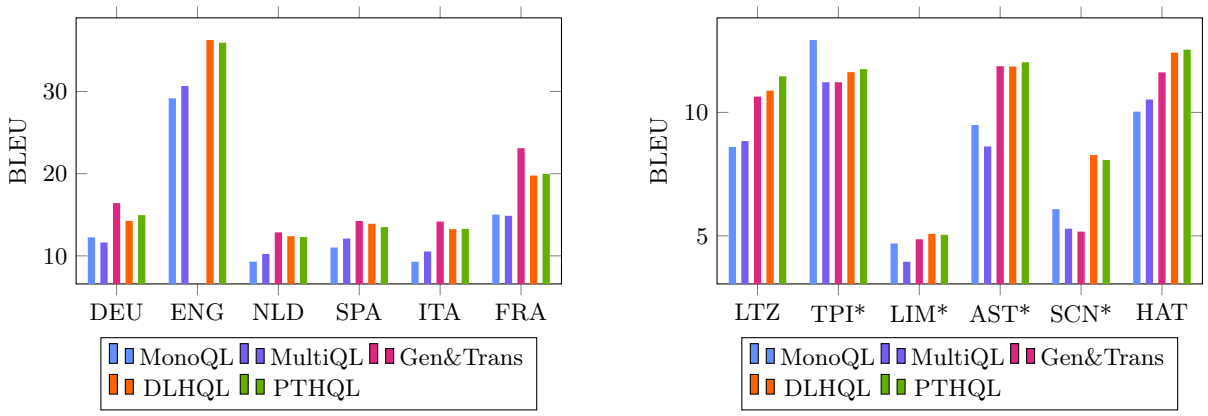
## 4.6 Results

Results obtained when generating from both silver and gold AMRs are reported. The proposed HQL approach is compared with previous works and baselines, and the results are examined on both high- and low-resource languages.

**HQL outperforms or is on par with mono and multilingual baselines (silver and gold AMRs).** On silver AMRs, HQL models are consistently better than mono and multilingual baselines, except for Tok Pisin (Figure 4.2, Table 4.3, Table 4.4, Table 4.5). This difference is statistically significant in most cases. The results on gold AMRs are more mixed. The proposed models outperform these baselines on Italian and German, but not on English and Spanish. This difference might be because both of the best-performing languages are among the most represented in the pre-training data of the base model.

**HQL outperforms the Gen&Trans Baseline on all LR languages.** While the Gen&Trans baseline outperforms the HQL models on most HR languages, the proposed approach outperforms the Gen&Trans models on all LR languages in terms of BLEU(Figure 4.2). These results show the benefits of HQL for LR languages. While MT yields low-quality texts, the stacked LoRA approach seems to enhance transfer. Similar results are seen on other metrics (Table 4.4, Table 4.5), where HQL comes ahead in most LR languages.

Notably, two of the languages previously unseen by the base model (Tok Pisin and Asturian) benefit from the transfer effect as they perform on par with LR languages from the base model's pre-training data. For Limburgish and Sicilian, the conjecture is that their lower scores result from the low quality of the machine translation, as evidenced by their particularly poor performance in the Gen&Trans baseline.




(a) High-Resource Languages

(b) Low-Resource Languages

FIG. 4.2: HQL BLEU on FLORES-200 test subset. *Languages unseen by the $mT5_{large}$ base model.

| | BLEU ↑ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Deu** | **Ltz** | **Eng** | **Tpi** | **Nld** | **Lim** | **Spa** | **ASt** | **Ita** | **Scn** | **Fra** | **Hat** |
| MonoQL | 12.2 | 8.6 | 29.2 | **12.9** | 9.3 | 4.7 | 11.0 | 9.5 | 9.3 | 6.1 | 15.0 | 10.0 |
| MultiQL | 11.6 | 8.8 | 30.7 | 11.2 | 10.2 | 4.0 | 12.1 | 8.6 | 10.5 | 5.9 | 14.9 | 10.5 |
| Gen&Trans* | **16.4** | 10.6 | — | 11.2 | **12.9** | 4.9 | **14.2** | 11.9 | **14.2** | 5.2 | **23.1** | 11.6 |
| DLHQL | 14.2 | 10.9 | **36.3** | 11.6 | 12.4 | **5.1** | 13.9 | 11.9 | 13.2 | **8.3** | 19.8 | 12.4 |
| PTHQL | 15.0 | **11.5** | 35.9 | 11.8 | 12.3 | 5.0 | 13.5 | **12.0** | 13.3 | 8.1 | 20.0 | **12.5** |

TAB. 4.3: HQL BLEU on FLORES-200 test subset. *English not included since no translation is needed.

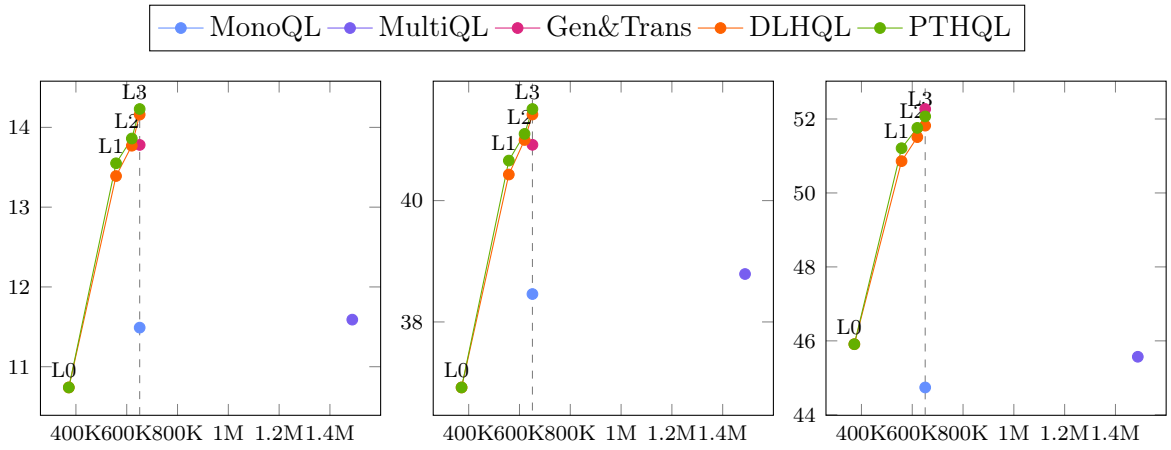
| | ChrF++ ↑ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Deu** | **Ltz** | **Eng** | **Tpi** | **Nld** | **Lim** | **Spa** | **ASt** | **Ita** | **Scn** | **Fra** | **Hat** |
| MonoQL | 39.2 | 37.1 | 58.5 | **38.4** | 36.8 | 30.8 | 37.1 | 34.3 | 37.0 | 32.5 | 42.2 | 37.6 |
| MultiQL | 39.8 | 37.0 | 60.7 | 36.6 | 38.4 | 30.2 | 39.1 | 32.8 | 37.9 | 32.8 | 42.3 | 37.8 |
| Gen&Trans* | **44.0** | **39.5** | — | 35.0 | **41.8** | 31.9 | **41.3** | **47.4** | **42.3** | 30.5 | **49.2** | 39.5 |
| DLHQL | 42.7 | 39.1 | **64.4** | 35.7 | 40.8 | **32.2** | **41.3** | 37.0 | 41.0 | **35.9** | 47.1 | **39.9** |
| PTHQL | 43.1 | 39.4 | 64.4 | 35.7 | 41.5 | 32.0 | 40.9 | 37.6 | 40.6 | 35.7 | 47.6 | **39.7** |

TAB. 4.4: HQL ChrF++ on FLORES-200 test subset. *English not included since no translation is needed.

| | BLEURT-20 ↑ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Deu** | **Ltz** | **Eng** | **Tpi** | **Nld** | **Lim** | **Spa** | **ASt** | **Ita** | **Scn** | **Fra** | **Hat** |
| MonoQL | 57.8 | 34.9 | 68.5 | 59.3 | 63.2 | 33.7 | 47.5 | 28.9 | 50.4 | 16.4 | 40.3 | 45.7 |
| MultiQL | 52.4 | 36.0 | 70.8 | 60.0 | 62.2 | 34.6 | 51.8 | 30.1 | 51.6 | 17.7 | 40.5 | 45.8 |
| Gen&Trans* | **64.9** | **43.2** | — | 59.7 | **64.7** | 37.6 | **59.3** | **38.2** | **63.2** | 21.8 | **56.8** | 49.6 |
| DLHQL | 61.2 | 42.0 | **74.5** | **60.6** | 55.5 | 37.6 | 58.5 | 37.4 | 60.7 | **22.3** | 53.3 | 51.4 |
| PTHQL | 61.4 | 42.2 | 74.3 | 59.9 | 53.5 | **38.0** | 58.8 | 38.1 | 61.1 | 22.2 | 54.3 | **51.4** |

TAB. 4.5: HQL BLEURT-20 on FLORES-200 test subset. *English not included since no translation is needed.

**HQL optimizes faster than the three baseline models and, on average, outperforms them all.** Figure 4.3 plots the average BLEU, Chrf++, and BLEURT-20 score for all 12 languages against the number of instances seen during training. Already at level L2, the HQL models outperform all three baselines (monolingual, multilingual, Gen&Trans ) on two metrics, despite having fewer total training instances. The graph also shows that each new level of the hierarchy improves performance.




(a) BLEU (b) ChrF++ (c) BLEURT-20

FIG. 4.3: HQL averaged scores vs training instances on FLORES-200 test subset. Average score across all 12 languages (Y axis) vs. total instances seen during training (X axis) for three metrics on the sub-set of FLORES-200 test data. HQL models include results on all the intermediary levels of the hierarchy.

**HQL performs on par with previous work (Gold AMRs).** Table 4.6 compares the results of the proposed models with previous works on gold AMRs. In HR Romance languages, the HQL approach outperforms all previous works in terms of BLEU score. In English, the score is close to the best-performing model. In German, the proposed model underperforms both Xu's and Martinez's approaches, possibly due to differences in training data size and the impact of multi-task learning.

| | **BLEU ↑** | | | | **ChrF++ ↑** | | | | **BLEURT-20 ↑** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Deu** | **Eng** | **Spa** | **Ita** | **Deu** | **Eng** | **Spa** | **Ita** | **Deu** | **Eng** | **Spa** | **Ita** |
| F&G | 15.3 | 24.9 | 21.7 | 19.8 | — | — | — | — | — | — | — | — |
| Ribeiro | 20.6 | — | 30.7 | 26.4 | 49.4 | — | 57.2 | 54.0 | — | — | — | — |
| Xu | **25.7** | — | 31.4 | 28.4 | — | — | — | — | — | — | — | — |
| Martinez | 23.2 | 44.8 | 34.6 | 29.0 | **55.8** | **73.4** | **64.0** | **60.7** | — | — | — | — |
| MonoQL | 18.2 | **49.2** | 38.6 | 22.7 | 45.4 | 71.1 | 61.8 | 50.9 | 60.7 | 78.1 | 68.1 | 63.1 |
| MultiQL | 19.8 | 42.9 | 34.1 | 27.2 | 47.7 | 69.7 | 60.1 | 54.7 | 57.7 | 77.6 | 66.6 | 65.9 |
| Gen&Trans* | **28.0** | — | **39.6** | **33.8** | 54.9 | — | 63.7 | 59.6 | **69.4** | — | **71.9** | **73.4** |
| DLHQL | 21.2 | 44.2 | 37.4 | 29.2 | 49.0 | 70.5 | 62.3 | 56.9 | 62.3 | 78.9 | 71.3 | 71.0 |
| PTHQL | 22.8 | 43.4 | 37.2 | 29.7 | 50.6 | 70.1 | 62.3 | 57.1 | 63.7 | **79.2** | 70.7 | 71.4 |

TAB. 4.6: HQL BLEU, ChrF++, and BLEURT-20 on LDC2020T07 test data. *English not included since no translation is needed.

**HQL performs well compared to previous works despite being trained on less data.** In previous works, *F&G, Ribeiro* and *Xu* trained on anywhere from 400k to 8.9M synthetic training pairs per language, in contrast, while the *Martinez* model is trained for up to 30 epochs on close to 55K monolingual instances. In contrast, the HQL models are trained on four epochs and fewer than 31K instances per language. Despite this, the proposed models come close to and in some cases, outperform those previous approaches, while also enabling support for LR languages.

**Distant vs. Close Languages.** Almost no significant difference between training on distant (DLHQL) and closely related (PTHQL) languages is observed. While this could confirm the observation by Meng and Monz (2024) that both are useful in inducing transfer and regularization, this could also be due to the restricted size of the training tree used. Due to computational constraints, the study was limited to a few languages, resulting in a substantial overlap of training data between the two hierarchies: 100% on L0 and L3, 50% on L1 and L2, for a total training overlap of 81%. To further evaluate the difference between these approaches, future studies could reduce the overlap by selecting a larger hierarchy, starting with a reduced number of instances and increasing their number as training progresses through the levels, or by using partial splits of data per language on the first levels.

Tables 4.7 and 4.8 on the next pages display generated text in a low-resource language (Tok Pisin) and its related high-resource language (English). The examples are taken from the 1st (best), 3rd (medium), and 5th (worst) quintiles, based on the average of the three automatic metrics used. Appendix B.1 provides similar examples for the other five low-resource languages and their five high-resource counterparts.

## 4.7 Conclusion

This chapter addressed the second research question by proposing Hierarchical QLoRA (HQL), a novel multilingual training strategy for AMR-to-Text generation that leverages phylogenetic information and parameter-efficient fine-tuning. The experiments across 12 languages, including both high- and low-resource settings, demonstrate that HQL consistently outperforms monolingual and multilingual baselines, especially in low-resource languages. Notably, it also surpasses a Generate-and-Translate pipeline in low-resource languages, despite relying heavily on machine-translated training data produced with the same translation model.

Two curriculum strategies were evaluated: one based on distant language groupings and the other on phylogenetic proximity. The differences in performance between the two were modest and, in most cases, not statistically significant, which supports previous studies on the regularizing effect of the first approach and the transfer learning of the second one. That being said, the phylogenetic hierarchy typically yielded better results, which opens the door to further testing on its suitability for distant language grouping.

These findings validate the use of structured phylogenetic information to inform multilingual training curricula and support the effectiveness of LoRA-based modular adaptation in large-scale, multilingual natural language generation (NLG) tasks.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(s2 / summarize-01<br>:ARG1 (s / situation<br>:location (c / country<br>:quant 1)<br>:mod (p / politics))<br>:ARG2 (a / advise-01)<br>:duration (b / brief)<br>:mod (m / mere))<br>**Generation (Tpi):**<br>Dispela em i wanpela liklik stori tasol bilong ol samting i kamap long wanpela kantri.<br>**Reference (Tpi):**<br>Ol 'advisory' em ol sotpela tok save long ol samting i kamap insait long politiks bilong wanpela kantri.<br>**Generation (Eng MT*):**<br>This is just a brief overview of what is happening in one country.<br>**Reference (Eng):**<br>Advisories are merely a brief summary of the political situation in one country. | 0.20 | 0.48 | 0.54 | 1st |
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Tpi):**<br>Tasol planti samting i luk olsem ol dinosaurus.<br>**Reference (Tpi):**<br>Tasol i gat planti samting long ol pisin we i luk wankain olsem ol dainaso yet.<br>**Generation (Eng MT*):**<br>But a lot of things look like dinosaurs.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.07 | 0.38 | 0.57 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Tpi):**<br>Ol i save kisim pe bilong skul bilong ol.<br>**Reference (Tpi):**<br>Planti taim bai i gat skul fi long enrol long ol dispela edukesen progrem.<br>**Generation (Eng MT*):**<br>They get paid for their education.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.04 | 0.18 | 0.49 | 5th |

Tab. 4.7: PTHQL Tok Pisin generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Tok Pisin generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Eng):**<br>But there are a lot of things about birds that still look like dinosaurs.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.81 | 0.90 | 0.85 | 1st |
| **Graph:**<br>(h / have-03<br>:ARG0 (c / country<br>:name (n / name<br>:op1 "Persian"))<br>:ARG1 (g / grammar<br>:ARG1-of (r / regular-02<br>:mod (m / most))<br>:ARG1-of (e / easy-05<br>:ARG2-of (r2 / relative-05))))<br>**Generation (Eng):**<br>Persian has mostly regular and relatively easy grammar.<br>**Reference (Eng):**<br>Persian has a relatively easy and mostly regular grammar. | 0.18 | 0.75 | 0.84 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Eng):**<br>Entry into these education programs is a typical tuition fee.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.08 | 0.44 | 0.77 | 5th |

Tab. 4.8: PTHQL English generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). **The graphs were automatically parsed and may contain errors.

# Chapter 5

# Referenceless Evaluation of Multilingual RDF-to-Text

**Contents**



While researching multilingual G2T, the limited number of informative neural evaluation metrics that work effectively in both high- and low-resource languages became evident. This chapter addresses the third research question: Can Natural Language Inference (NLI) be used as the base to develop a referenceless multilingual evaluation metric for multiple facets of semantic faithfulness in RDF-to-Text generation across high- and low-resource languages?

## 5.1 Introduction

As discussed in Chapter 1, a key constraint on the G2T task is that generation should be *semantically faithful*, meaning that the generated text should express only the content represented by the input graph and represent it completely.

While G2T generation models have steadily improved over the years both in terms of performance and range of target languages they can handle (Gardent et al., 2017; Castro Ferreira et al., 2020; Cripwell et al., 2023), recent results indicate that semantic faithfulness is still an issue since the generated texts can either contain information not present in the input (Additions) or, conversely, fail to express all the information present in the input (Omissions). These issues are particularly prevalent when generating into under-resourced languages (Cripwell et al., 2023) or out-of-domain topics (Nikiforovskaya and Gardent, 2024).

In this chapter, a novel framework for evaluating G2T Models is proposed, aiming to support the development of multilingual, semantically faithful G2T models. In particular, the following contributions are made:

1) A new referenceless multilingual metric that quantifies how much a model under- (omissions) or over- (additions) generates. This metric provides three scores: precision, recall, and F1. Intuitively, the graph acts as a reference. Hence, precision is the ratio between correct information in the text and total information in the text (how much of the generated text is correct), recall is the ratio between correct information in the text and information in the input graph (how much of the input graph does the text convey), and F1 is their harmonic mean.

2) A methodology for creating the training data necessary to train this metric.

3) Tests on both high (English, Russian) and low (Breton, Irish, Maltese, Welsh, Xhosa) resource languages and compute correlation with both existing reference-based metrics and human judgments. The tests show that the correlation with reference-based metrics is fair to moderate, indicating that the proposed metric, although referenceless, can be used to a certain extent in place of reference-based metrics, particularly when references are unavailable. When comparing with human judgments, results show that correlation with the proposed metric outperforms the correlation obtained on the same data by other existing referenceless metrics developed for English G2T, like Data-QuestEval Rebuffel et al. (2021) and FactSpotter Zhang et al. (2023).

## 5.2 Method

To learn this metric, an existing multilingual Natural Language Inference (NLI) model is further fine-tuned by adjusting its classification head to work as a regression model instead. This fine-tuning is performed on synthetic data created to capture various combinations of precision and recall using Binary Cross-Entropy (BCE) loss. The intuition is the following.

Given a premise and a hypothesis, NLI models predict if the hypothesis is entailed, neutral, or contradicted by the premise. For precision, the model checks if the text is entailed by the graph (how much of the text can be inferred from the graph). For recall, the model checks if the graph is entailed by the text (how much of the graph content can be inferred from the text).

Since there is no interest in the three classes from the NLI head, only in the degree of the entailment between the premise and the hypothesis. The NLI classifier is further fine-tuned as a regression model by focusing solely on the entailment weights from the classification head, rather than the three existing output classes. The model is trained simultaneously for precision and recall by swapping the order of the graph and text and targeting the respective score.

The F1 score is computed as usual (Equation 5.1) by taking the harmonic mean of precision (p) and recall (r). This score functions as a high-level proxy for semantic faithfulness: the higher the F1 score, the higher the semantic similarity between the Graph and the Text.

$$F_1 = 2\,\frac{p \cdot r}{p + r} \tag{5.1}$$

## 5.3 Data

### 5.3.1 Training Data

The aim is to generate a training dataset of (graph, text, precision, recall) quadruples that exhibit a balanced and diverse distribution of precision and recall combinations.

**True Graph Collection:** The data creation process starts with the English WebNLG V3.0 dataset (Castro Ferreira et al., 2020)[48]. This dataset contains aligned (graph $g_i$, text $t_i$) pairs. Graph $g_i$ can also be referred to as $g_{t_i}$, meaning it is the graph aligned with $t_i$. In the dataset, the graphs were extracted from DBPedia,[49], and the texts were either automatically lexicalized or mined from Wikipedia[50] before being aligned with each other by human annotators. Since the graph and text are aligned, the pair has precision and recall scores equal to 1, forming a quadruple (graph $g_{t_i}$, text $t_i$, precision=1, recall=1).

Variations of these original quadruples can be created with diverse precision and recall scores by finding pairs (graph $g_j$, text $t_i$) with different levels of information overlap ($o$). To do so, keep the text static and change the graph, as it is much easier to work with and manipulate data in a graphical representation. For example, measuring $o$ is much easier when both elements are in graph representation since it is possible to compute the intersection between both sets of triples ($o = |g_j \cap g_{t_i}|$). Because of that, for most of the creation process, work is done with (graph $g_j$, graph $g_{t_i}$) pairs instead of (graph $g_j$, text $t_i$) pairs. At the end of the process, $g_{t_i}$ is substituted with the original text $t_i$.

Starting with the graph $g_{t_i}$, to obtain a variation quadruples with precision $p$ and recall $r$, find a new $g_j$ such that the following equations are true:

- $o/|g_{t_i}| = p$
- $o/|g_j| = r$

At first, such a $g_j$ can be searched for in the list of all original graphs from WebNLG and all its subgraphs. If finding a matching graph is impossible, a synthetic one that satisfies the criteria can be created.

[48] https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0
[49] https://www.dbpedia.org/
[50] https://www.wikipedia.org/

**Synthetic Graph Creation:** Given a graph $g_{t_i}$, a synthetic graph $g_j$ with precision $p$ and recall $r$ can be created by first taking $o$ triples from $g_{t_i}$ and then adding external triples to $g_j$ until $o/|g_j| = r$.

The external triples can be procured by selecting a triple from some graph $g_k$ that has no overlap with $g_{t_i}$ or by corrupting real triples from $g_{t_i}$ so that the information they represent does not match the original graph.

Corrupting a real triple can be achieved by swapping the order of the elements in the triple or substituting some or all of its elements with incorrect values. When doing so, logical substitutions are used. For example, to corrupt the triple (Alan Bean | birthPlace | Wheeler, Texas), the object *Wheeler, Texas* can be swapped with a different value. In such a case, a value that can be paired with the property *birthPlace* is selected, like *Miami, Florida*, instead of a random value like *1932-03-15*.

**Multilingual Text Generation:** Once a balanced English dataset has been created, it can be extended to other languages by machine translating the text. NLLB-200-3.3B (NLLB Team et al., 2022)[51] is used to translate into five languages: Irish, Maltese, Russian, Welsh, and Xhosa. To reduce the noise introduced by machine translation, these translations are filtered following two criteria: Language Identification score via GlotLID Kargaran et al. (2023)[52] and LID218e (NLLB Team et al., 2022)[53], and semantic similarity score via LaBSE (Feng et al., 2022)[54].

The resulting dataset contains approximately 1.77 million quadruples (graph, text, precision, recall) evenly distributed across six languages. Figure 5.1 shows the distribution of precision and recall scores in the final dataset. Appendix C.2

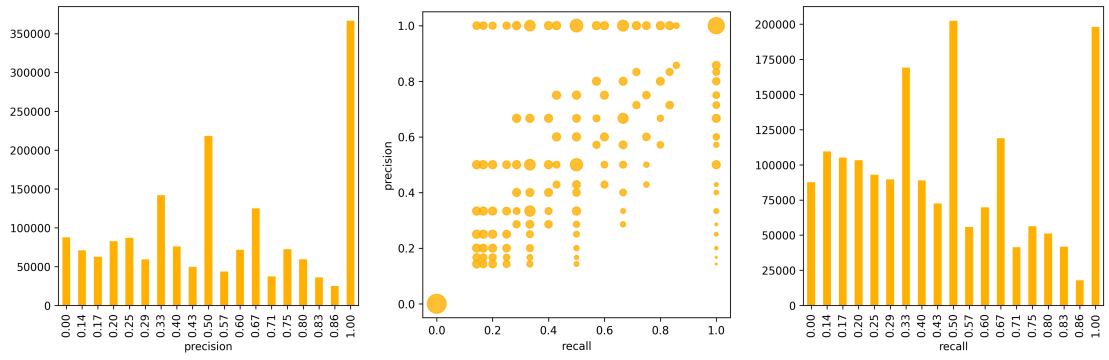



FIG. 5.1: Referenceless metric synthetic dataset by number of samples according to precision and recall.

### 5.3.2 Test Data

Testing relies exclusively on text generated by real models submitted to different G2T challenges to evaluate the proposed models. Table 5.1 summarizes the datasets used to perform evaluation.

[51] https://huggingface.co/facebook/nllb-200-3.3B
[52] https://huggingface.co/cis-lmu/glotlid
[53] https://huggingface.co/facebook/fasttext-language-identification
[54] https://huggingface.co/sentence-transformers/LaBSE

| Dataset | Graphs | Texts | Languages | Relevant Annotations |
|---|---|---|---|---|
| 7L-Auto | 4461 | 143 838 | Bre, Cym, Eng, Gle, Mlt, Rus, Xho | BLEU, ChrF++, TER, BERTScore, SBERT |
| 2017 | 223 | 2 230 | Eng | Semantics |
| 2020 | 288 | 3 905 | Eng, Rus | Relevance*Correctness(p), Data Coverage(r) |
| 2023 | 200 | 1 700 | Cym, Gle, Rus, Mlt | Omissions(r), Additions(r) |
| 4L-RP-Human | 181 | 200 | Cym, Eng, Mlt, Rus | Precision(p) and Recall(r) |

TAB. 5.1: Referenceless metric test datasets used for correlation studies. When a relevant annotation is adjacent to precision (p) or recall (r) that is indicated, otherwise the annotation is consider adjacent to the more general F1 score.

**7L-Auto:** This dataset consists of all graphs from the WebNLG test data[55] and all the texts generated from these graphs by participant systems of the WebNLG 2017, 2020, and 2023 Shared Tasks, as well as the different models trained by Meyer and Buys (2024). The models used to generate the texts encompass a range of approaches, including grammar-based and template-based methods, statistical machine translation models, neural models trained from scratch, and fine-tuned pre-trained models, which cover a broad spectrum of errors and quality levels. Texts are generated in English (Eng), Russian (Rus), Breton (Bre), Irish (Gle), Maltese (Mlt), Welsh (Cym), and Xhosa (Xho).

**WebNLG 2017.** The human annotations for this challenge (Shimorina et al., 2018) consist of 223 graphs lexicalized in English by nine different NLG systems, plus the human-written references. The generations were scored on a 3-point Likert scale across three criteria: fluency, grammar, and semantics. For this study, the focus was only on the semantics annotation, which was defined as follows:

- Semantics: *Does the text correctly represent the meaning in the data?*

**WebNLG 2020.** The human annotations for this challenge (Castro Ferreira et al., 2020) consist of 178 graphs lexicalized in English by 16 different NLG systems and 110 graphs lexicalized in Russian (Rus) by seven different NLG systems; additionally, both include their human-written references. The generations were scored on a 0-100 scale across five criteria: text structure, fluency, relevance, correctness, and data coverage. For this study, the focus was on the last three, which were defined as follows:

- Relevance: *Does the text describe only such predicates (with related subjects and objects), which are found in the data?*
- Correctness: *When describing predicates which are found in the data, does the text correctly mention the objects and adequately introduce the subject for this specific predicate?*
- Data Coverage: *Does the text include descriptions of all predicates presented in the data?*

[55] Specifically, the following graphs are used: from the WebNLG 2017 test set (1,862 graphs), from the WebNLG 2020 test set for English (1,779), from the WebNLG 2020 test set for Russian that are not present in the English test set (732) and from the Xhosa data sets that are not in any of the other datasets (88).

**WebNLG 2023.** The human annotations for this challenge (Cripwell et al., 2023) consist of 100 graphs lexicalized in Irish by 4 NLG systems, Maltese by 3 NLG systems, Welsh by 3 NLG systems, and another 100 graphs lexicalized in Russian by 3 NLG systems. Additionally, all of them included their human-written references. The generations were scored across four different criteria: fluency, absence of unnecessary repetition, absence of additions, and absence of omissions. The first was on a 5-point Likert scale, and the other three had binary Yes/No labels. For this study, the focus was exclusively on the absence of additions and the absence of omissions, which were defined as follows:

- Absence of Additions: *Looking at the Text, is all of its content expressed in the Data expression? (Allow duplication of content.)*
- Absence of Omissions: *Looking at each element of the Data expression in turn, does the Text express all the information in all elements in full (allow synonyms and aggregation)?*

**4L-RP-Human.** While WebNLG's existing human judgments can, to a certain extent, be used as proxies for precision, recall, and F1, none of them were collected to measure these values specifically. To address this, a new dataset of human judgments called 4L-RP-Human was created, with graph-text pairs extracted from the 7L-Auto dataset. It contains 50 graph-text pairs per language for four languages (English, Maltese, Russian, and Welsh) with a balanced distribution of precision and recall scores by the best-performing model. Fine-grained human annotations were obtained for precision and recall of this subset to test how the proposed models correlate with human judgments that specifically target these properties. The human annotators were provided with a text and a graph in table format and were asked to answer, using a scale of 1 to 5 (None, Few, Half, Most, All), the following questions:

- Precision: *How many Triples from the text can you find in the Table?*
- Recall: *How many Triples from the Table can you find in the Text?*

The annotators were native speakers of the target language who were proficient in English. They were hired via Prolific [56] and paid 10£/h. To measure inter-annotator agreement, the Fleiss' Kappa (Fleiss, 1971) was used. Appendix C.2 provides more detail on the annotation process.

## 5.4 Experiments

### 5.4.1 Training Process

The proposed models are trained by fine-tuning mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (NB) on the new synthetic dataset to learn the metric. They are fine-tuned as a regression model by targeting only the entailment weights of the classification head and training simultaneously for both precision and recall. For the precision score, the graph serves as the premise, and the text serves as the hypothesis. For the recall score, the text serves as the premise, and the graph serves as the hypothesis.

[56] https://www.prolific.com/

### 5.4.2 Models

#### Baselines

*Data-QuestEval(DQE)* is referenceless model by Rebuffel et al. (2021). It utilizes question generation and question answering to assess semantic Faithfulness. Its main limitations are being fine-tuned only in English, the lengthy processing time of text generation, and the risk of generating questions and answers unrelated to the input.

*FactSpotter(FS)* is the latest model by Zhang et al. (2023)[57]. Its main limitations are that it is fine-tuned only in English and produces only a recall-oriented score.

*NLI Base (NB)* consists on following the process from Dušek and Kasner (2020) with a different (multilingual) off-the-shelf NLI model (Laurer et al., 2022)[58] instead of an English one. Its main limitations are being unfamiliar with the RDF graph format and not having tested all target languages.

#### *Proposed Models*

**Multilingual Full Fine-Tuning (MultiFF):** Full fine-tuning the NLI Base model on all target languages simultaneously.

**Multilingual LoRA (MultiLR):** LoRA fine-tuning on top of the NLI Base model on all target languages simultaneously.

**Monolingual LoRA (MonoLR):** LoRA fine-tuning on top of the NLI Base model on each target language individually.

Table 5.2 provides details on the hyperparameters used to train the proposed models.

| Parameter | MultiFF | MultiLR | MonoLR |
|---|---|---|---|
| Training Hardware | 1 32GB V100 | 1 32GB V100 | 1 32GB V100 |
| Training Instances | ∼ 3 544 994 | ∼ 3 544 994 | ∼ 590 832 |
| Training Epochs | 1 | 1 | 1 |
| Training Time | ∼ 7h | ∼ 11h | ∼ 2h |
| Warmup Steps | 10% | 10% | 10% |
| Scheduler | WarmupLinear | WarmupLinear | WarmupLinear |
| Optimizer | AdamW | AdamW | AdamW |
| Learning Rate | 2e-5 | 2e-5 | 2e-5 |
| Loss Function | BCELoss | BCELoss | BCELoss |
| Rank | N/A | 32 | 32 |
| Total parameters | 278 811 651 | 283 507 299 | 283 507 299 |
| Trained parameters | 278 811 651 | 5 382 240 | 5 382 240 |

Tab. 5.2: Referenceless metric hyperparameters.

[57] Inria-CEDAR/FactSpotter-DeBERTaV3-Base
[58] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

## 5.5 Evaluation

The metric is evaluated using correlation with human judgments (in 6 languages) and with automatic metrics (in 7 languages). Additional results are reported on a graph-text retrieval accuracy experiment (7 languages).

### 5.5.1 Correlation with Automatic Metrics

For this experiment the **7L-Auto** dataset is used by computing the Spearman's Correlation ($\rho$) of the baselines and the proposed models with five reference-based metrics: BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), TER (Snover et al., 2006), BERTScore (Zhang et al., 2020b), and SBERT similarity (Reimers and Gurevych, 2019). For TER, its inverse score ($\neg$TER $=$ $1 -$ TER) is reported for a more uniform reading and display. Since none of these metrics specifically targets semantic precision or semantic recall, they are compared against the semantic F1 score. The intuition is that a good correlation with these standard metrics suggests that, in the absence of ground truth to use them, the proposed metrics can serve as a proxy for them.

### 5.5.2 Correlation with Human Judgments

First, the human evaluations from all three WebNLG shared tasks were used. Since none of those have direct fine-grained scores for semantic precision and semantic recall, they were compared as follows:

**WebNLG 2017:** The only closely related annotation from this dataset is semantics. Since it does not specify the type of error (additions, omission, etc.), it was compared against the semantic F1 obtained by computing the harmonic mean of semantic precision and semantic recall.

**WebNLG 2020:** semantic recall is compared with the annotations for data coverage since both are very similar. Semantic precision is compared with the product of the annotations of relevance and correctness since both are related to precision. However, based on the annotation definitions, correctness only refers to already relevant triples. For F1, the harmonic mean of the scores for precision and recall was used.

**WebNLG 2023:** semantic recall is compared with the annotations for the absence of omissions, despite the lack of granularity in that binary label. Similarly, semantic precision is compared with the annotations for the absence of additions, keeping in mind the binary nature of those labels. For F1, the harmonic mean of the scores for precision and recall was used.

Additionally, the newly collected **4L-RP-Human** dataset is used, since it has dedicated semantic precision and semantic recall scores, to evaluate the best-performing model.

### 5.5.3 Retrieval Accuracy

Finally, how well the scoring of various models discerns good pairings from bad ones using a retrieval method was evaluated: Does the metric give higher scores to pairs where the graph and the generation are equivalent than to pairs where they are not?

Given a subset of 100 graph-text pairs randomly selected from the WebNLG dataset for each target language and their corresponding human references, the F1 score is computed with the

proposed models and the score produced by each of the baselines for each of the 10K possible (graph, text) combinations. Then, Accuracy at 1 (A@1), the proportion of cases where the highest score is assigned to the correct graph-text pair, is computed. The size of this subset is limited given the computational demands of scoring all 10 000 possible combinations of graphs and texts with the proposed cross-encoder approach.

## 5.6 Results

### 5.6.1 Correlation with Automatic Metrics

Table 5.3 shows the Spearman's Correlation ($\rho$ ) between the various referenceless metrics evaluated and the reference-based automatic metrics on the 7L-Auto dataset (Breton, English, Irish, Maltese, Russian, Welsh and Xhosa WebNLG generations) in terms of precision, recall, and F1 score. Figure 5.2 summarizes the F1 results.

| | Breton | | | | | English | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | F1 | | | | | F1 | | |
| | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** |
| DQE | 0.24 | 0.30 | **0.18** | 0.31 | 0.35 | **0.51** | **0.60** | 0.50 | **0.62** | **0.68** |
| FS | 0.25 | 0.28 | — | 0.29 | 0.32 | **0.51** | **0.60** | 0.46 | 0.61 | 0.67 |
| NB | — | — | — | -0.09 | — | -0.27 | -0.30 | -0.39 | -0.36 | -0.30 |
| MultiFF | 0.37 | 0.41 | 0.12 | 0.39 | 0.34 | 0.36 | 0.47 | 0.41 | 0.48 | 0.54 |
| MultiLR | 0.43 | **0.52** | **0.18** | 0.47 | **0.41** | 0.40 | 0.53 | 0.47 | 0.54 | 0.60 |
| MonoLR | **0.45** | 0.50 | 0.16 | **0.49** | 0.39 | 0.44 | 0.58 | **0.53** | 0.61 | 0.67 |

| | Irish | | | | | Maltese | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | F1 | | | | | F1 | | |
| | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** |
| DQE | 0.23 | 0.21 | 0.12 | 0.17 | 0.23 | 0.54 | 0.60 | 0.37 | 0.55 | 0.60 |
| FS | 0.29 | 0.31 | 0.17 | 0.28 | **0.33** | 0.60 | 0.66 | 0.38 | 0.62 | 0.62 |
| NB | -0.12 | -0.11 | -0.18 | -0.21 | -0.07 | 0.08 | 0.10 | — | 0.02 | 0.14 |
| MultiFF | 0.28 | 0.29 | 0.14 | 0.29 | 0.21 | 0.70 | 0.78 | 0.46 | 0.72 | 0.74 |
| MultiLR | 0.38 | 0.40 | 0.20 | 0.39 | 0.27 | 0.72 | 0.80 | **0.49** | 0.74 | **0.78** |
| MonoLR | **0.40** | **0.41** | **0.22** | **0.41** | 0.29 | **0.76** | **0.84** | **0.49** | **0.77** | **0.78** |

| | Russian | | | | | Welsh | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | F1 | | | | | F1 | | |
| | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** |
| DQE | -0.05 | -0.07 | -0.03 | -0.08 | -0.08 | 0.34 | 0.37 | 0.29 | 0.37 | 0.47 |
| FS | -0.02 | — | 0.02 | — | 0.04 | 0.36 | 0.40 | 0.29 | 0.41 | 0.47 |
| NB | -0.07 | -0.12 | -0.22 | -0.22 | -0.07 | -0.09 | -0.09 | -0.16 | -0.19 | -0.07 |
| MultiFF | 0.13 | 0.19 | 0.11 | 0.18 | 0.23 | 0.46 | 0.49 | 0.32 | 0.49 | 0.44 |
| MultiLR | **0.26** | **0.37** | **0.29** | **0.39** | **0.35** | **0.54** | **0.59** | **0.39** | **0.59** | **0.51** |
| MonoLR | 0.25 | 0.36 | 0.28 | 0.38 | 0.34 | 0.53 | 0.58 | 0.37 | **0.59** | **0.51** |

| | Xhosa | | | | |
|---|---|---|---|---|---|
| | | | F1 | | |
| | **BLEU ↑** | **ChrF++ ↑** | **¬TER ↑** | **BERTScore ↑** | **SBERT ↑** |
| DQE | -0.10 | -0.04 | -0.14 | -0.11 | -0.12 |
| FS | 0.19 | 0.18 | 0.18 | 0.11 | 0.13 |
| NB | -0.25 | -0.27 | -0.30 | -0.24 | -0.26 |
| MultiFF | — | -0.05 | -0.11 | — | -0.05 |
| MultiLR | 0.19 | 0.32 | 0.15 | 0.22 | 0.21 |
| MonoLR | **0.22** | **0.34** | **0.19** | **0.26** | **0.25** |

TAB. 5.3: Referenceless metric correlation with automatic metrics. Spearman's Correlation ($\rho$ ) of the F1 score from different automatic metrics against classic reference-based metrics on the 7L-Auto dataset. Only results with a p-value under 0.05 are reported. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with its closest language (Welsh).*

**Fine-tuning matters.** Models that have been fine-tuned for the task (including the FS baseline) show a positive correlation across metrics and languages. In contrast, the NB baseline

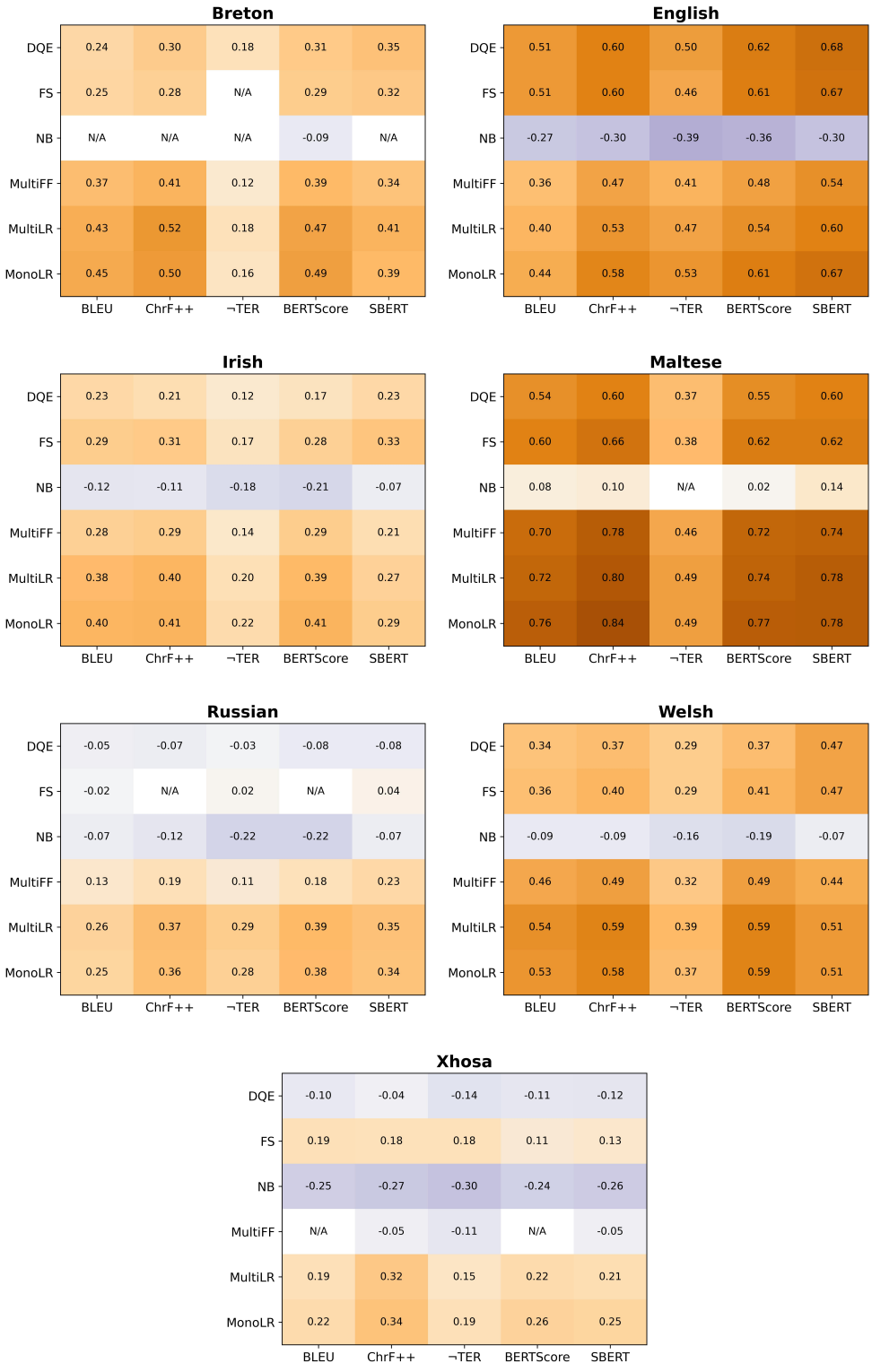



FIG. 5.2: Referenceless metric correlation with automatic metrics. Spearman's Correlation ($\rho$ ) between referenceless and reference-based metrics on the 7L-Auto dataset. Only results with a p-value under 0.05 are reported. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with its closest language (Welsh).*

produces either negative or non-significant correlations. These results highlight the limitations of using off-the-shelf models as proposed by Dušek and Kasner (2020) and underscore the importance of task-specific fine-tuning.

**Strong performance in English.** While trained on multilingual data, the proposed models almost match the performance of metrics trained on English only (DQE, FS, NB). Interestingly, the gap is smallest for semantic-based metrics (SBERT, BERTScore), suggesting that the metrics are good at capturing paraphrases.

**Good performance in other languages.** The fine-tuned models, especially the small Monolingual LoRA version, outperform all three baselines in all the other languages. These results demonstrate the effectiveness of the proposed approach despite fine-tuning on synthetic, non-gold data.

### 5.6.2 Correlation with Human Judgments

***Indirect Precision and Recall: The WebNLG Shared Task***

Figure 5.3 shows the Root Mean Squared Error (RMSE) and Spearman's correlation ($\rho$) of the F1 score from different automatic metrics against the WebNLG 2017, 2020, and 2023 human annotations. Exact numbers, as well as a breakdown by precision and recall (when possible), can be found in Table 5.4, Table 5.5, and Table 5.6

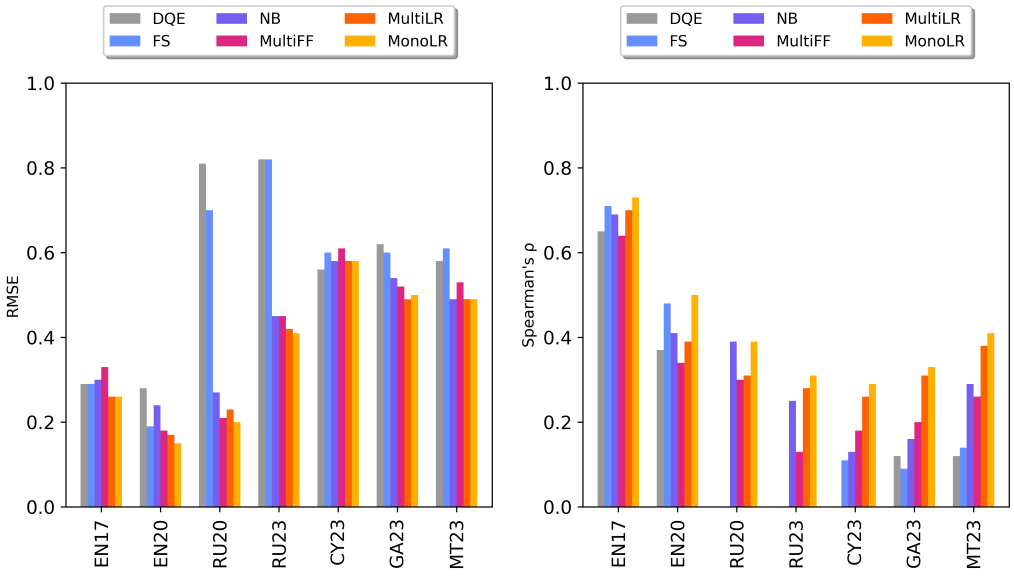



FIG. 5.3: Referenceless metric error and correlation with WebNLG human annotations. In particular, Root Mean Squared Error (RMSE) and Spearman's correlation ($\rho$ ) of the F1 score from different automatic metrics against the closest approximate human annotations from *WebNLG 2017, 2020, and 2023* (See Table 5.1 for more details). For Spearman's correlation scores, only results with a p-value under 0.05 are reported.

| | English | |
|---|---|---|
| | F1 | |
| | RMSE ↓ | $\rho$ ↑ |
| DQE | 0.29 | 0.65 |
| FS | 0.29 | 0.71 |
| NB | 0.30 | 0.69 |
| MultiFF | 0.33 | 0.64 |
| MutiLR | **0.26** | 0.70 |
| MonoLR | **0.26** | **0.73** |

TAB. 5.4: Referenceless metric error and correlation with WebNLG 2017 human annotations. Root Mean Squared Error (RMSE) and Spearman's correlation ($\rho$ ) of the F1 score from different automatic metrics with the *English WebNLG 2017* human annotations. Spearman's correlation is only reported where the p-value is less than 0.05.

| | English | | | | | | Russian | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | F1 | | Precision | | Recall | | F1 | |
| | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ |
| DQE | 0.27 | 0.37 | 0.32 | 0.32 | 0.28 | 0.37 | 0.79 | — | 0.84 | — | 0.81 | — |
| FS | 0.21 | 0.35 | 0.18 | 0.45 | 0.19 | 0.48 | 0.67 | — | 0.79 | — | 0.70 | — |
| NB | 0.28 | 0.28 | 0.16 | 0.43 | 0.24 | 0.41 | 0.30 | 0.21 | 0.20 | 0.39 | 0.27 | 0.39 |
| MultiFF | 0.22 | 0.22 | 0.15 | 0.38 | 0.18 | 0.34 | 0.25 | 0.14 | **0.14** | 0.42 | 0.21 | 0.30 |
| MultiLR | 0.22 | 0.36 | 0.20 | 0.37 | 0.17 | 0.39 | 0.26 | 0.19 | 0.22 | 0.36 | 0.23 | 0.31 |
| MonoLR | **0.20** | **0.44** | **0.14** | **0.47** | **0.15** | **0.50** | **0.24** | **0.25** | 0.16 | **0.44** | **0.20** | **0.39** |

TAB. 5.5: Referenceless metric error and correlation with WebNLG 2020 human annotations. Root Mean Squared Error (RMSE) and Spearman's correlation ($\rho$ ) of the precision, recall, and F1 score from different automatic metrics with the *English and Russian WebNLG 2020* human annotations. Spearman's correlation is only reported where the p-value is less than 0.05.

| | Irish | | | | | | Maltese | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | F1 | | Precision | | Recall | | F1 | |
| | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ |
| DQE | 0.65 | 0.11 | 0.64 | 0.09 | 0.62 | 0.12 | 0.60 | 0.14 | 0.59 | 0.10 | 0.58 | 0.12 |
| FS | 0.62 | — | 0.57 | 0.13 | 0.60 | 0.09 | 0.66 | 0.13 | 0.52 | 0.17 | 0.61 | 0.14 |
| NB | 0.52 | 0.14 | 0.46 | 0.22 | 0.54 | 0.16 | 0.48 | 0.20 | 0.47 | 0.37 | 0.49 | 0.29 |
| MultiFF | 0.48 | 0.14 | 0.49 | 0.29 | 0.52 | 0.20 | 0.46 | 0.14 | 0.52 | 0.32 | 0.53 | 0.26 |
| MultiLR | 0.47 | 0.18 | **0.45** | 0.35 | **0.49** | 0.31 | 0.44 | 0.26 | 0.46 | 0.37 | **0.49** | 0.38 |
| MonoLR | **0.46** | **0.21** | **0.45** | **0.37** | 0.50 | **0.33** | **0.43** | **0.30** | **0.45** | **0.43** | **0.49** | **0.41** |

| | Russian | | | | | | Welsh | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | F1 | | Precision | | Recall | | F1 | |
| | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ | RMSE ↓ | $\rho$ ↑ |
| DQE | 0.85 | — | 0.87 | — | 0.82 | — | 0.61 | — | 0.58 | 0.11 | 0.56 | — |
| FS | 0.76 | — | 0.86 | — | 0.82 | — | 0.69 | — | 0.55 | 0.18 | 0.60 | 0.11 |
| NB | 0.45 | 0.17 | 0.34 | 0.28 | 0.45 | 0.25 | 0.54 | 0.13 | **0.53** | 0.24 | **0.58** | 0.13 |
| MultiFF | 0.42 | — | 0.36 | 0.22 | 0.45 | 0.13 | 0.51 | — | 0.58 | 0.26 | 0.61 | 0.18 |
| MultiLR | 0.39 | 0.24 | 0.35 | 0.28 | 0.42 | 0.28 | **0.49** | 0.18 | 0.54 | 0.29 | **0.58** | 0.26 |
| MonoLR | **0.38** | **0.25** | **0.32** | **0.33** | **0.41** | **0.31** | **0.49** | **0.21** | 0.54 | **0.34** | **0.58** | **0.29** |

TAB. 5.6: Referenceless metric error and correlation with WebNLG 2023 human annotations. Root Mean Squared Error (RMSE) and Spearman's correlation ($\rho$ ) of the precision, recall, and F1 score from different automatic metrics with the *Irish, Maltese, Russian, and Welsh WebNLG 2023* human annotations. Spearman's correlation is only reported where the p-value is less than 0.05. *NB: Original human annotations are binary, which might explain the high RMSE.*

**Correlation is highest for 2017 Data:** The human judgments collected during the first WebNLG campaigns do not directly target precision and recall: WebNLG 2017 targets semantic faithfulness, WebNLG 2020 targets three semantic criteria related to but not identical to precision and recall, and WebNLG 2023 focuses on omissions and additions but only return a binary score no matter how much omission/addition occurs in the generated text. The available human scores were used to approximate an F1 score and compute the correlation between these derived F1 scores and each evaluated metric. The hypothesis is that the higher correlation obtained for the 2017 data results from the fact that, for that campaign, the single score provided by the human evaluation is more directly related to the unique score provided by the baseline metrics and to the F1 score. Conversely, the lower correlation scores obtained by the proposed models on the WebNLG 2020 and 2023 datasets are likely due to the need to "reconstruct" an F1 score from the human judgments provided in these datasets.

**An improvement over the state-of-the-art.** In English, the MonoLR model outperforms the three baselines despite these being optimized for the language. For the other four languages, the gap with these monolingual metrics is particularly pronounced. Surprisingly, for Russian, the NB model is on par with the MonoLR model. These results highlight the impact of using a multilingual model as the base model. However, the low results of the NB model on the other languages show that using NLI only, without fine-tuning on task-specific data, does not always suffice.

### *Direct Precision and Recall Evaluation: The 4L-RP-Human*

Table 5.7 reports correlation results when comparing the precision, recall, and F1 scores predicted by the best-performing model with corresponding human judgments (4L-RP-Human dataset). These results show a strong correlation for all three metrics for English, Russian, and Welsh, and a moderate one for Maltese, demonstrating the effectiveness of the proposed approach in capturing omissions (recall), additions (precision), and semantic faithfulness ($F_1$).

| **Language** | **Annotators** | **Precision** | | **Recall** | | **F1** |
|---|---|---|---|---|---|---|
| | | Fleiss $\kappa$ | $\rho$ | Fleiss $\kappa$ | $\rho$ | $\rho$ |
| English | 4 | 0.47 | 0.68 | 0.47 | 0.63 | 0.70 |
| Maltese | 3 | 0.29 | 0.38 | 0.49 | 0.30 | 0.47 |
| Russian | 2 | 0.32 | 0.63 | 0.39 | 0.52 | 0.67 |
| Welsh | 4 | 0.37 | 0.60 | 0.50 | 0.81 | 0.70 |

Tab. 5.7: Referenceless metric human evaluation on 4L-RP-Human. In particular, Fleiss' $\kappa$ of precision and recall human annotations as well as the Spearman's correlation ($\rho$) of their average compared to the MonoLR model.

Tables 5.8 and 5.9 display evaluation examples in a low-resource language (Welsh) and a high-resource language (English). The examples are taken from the 1st (best), 3rd (medium), and 5th (worst) quintiles, based on the accuracy of MonoLR compared to the human annotation. Appendix C.3 provides similar examples for the other languages from 4L-RP-Human (Maltese and Russian).

| **Sample** | **Precision** | | **Recall** | | **Q** |
|---|---|---|---|---|---|
| | **H** | **mLR** | **H** | **mLR** | |
| **Graph:**<br>McVeagh of the South Seas \| imdbId \| 0004319<br>McVeagh of the South Seas \| director \| Cyril Bruce<br>McVeagh of the South Seas \| director \| Harry Carey (actor born 1878)<br>McVeagh of the South Seas \| starring \| Harry Carey (actor born 1878)<br>McVeagh of the South Seas \| writer \| Harry Carey (actor born 1878)<br>**Evaluated Generation (Cym):**<br>Ysgrifennodd Harry Carey (a anwyd yn 1878) McVeagh of the South Seas a cyfarwyddodd Cyril Bruce. Mae gan y ddata IMDb 0004319.<br>**Evaluated Generation (Eng MT*):**<br>Harry Carey (born 1878) wrote McVeagh of the South Seas and Cyril Bruce directed. The IMDb data has 0004319. | 0.88 | 0.96 | 0.69 | 0.65 | 1st |
| **Graph:**<br>University of Burgundy \| campus \| Dijon<br>Dijon \| country \| France<br>**Evaluated Generation (Cym):**<br>Mae Prifysgol Burgundaidd yn cael ei leoli yn Dijon, Ffrainc.<br>**Evaluated Generation (Eng MT*):**<br>Burgundian University is located in Dijon, France. | 1.00 | 0.76 | 0.94 | 0.85 | 3rd |
| **Graph:**<br>Bionico \| dishVariation \| Honey Bionico \| country \| Mexico<br>**Evaluated Generation (Cym):**<br>Mae Bionico yn amrywiad dysgl o Fecsico sy'n cynnwys mêl.<br>**Evaluated Generation (Eng MT*):**<br>Bionico is a variation of a Mexican dish that includes honey. | 1.00 | 0.36 | 1.00 | 0.34 | 5th |

Tab. 5.8: Referenceless metric evaluation on Welsh examples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the accuracy of MonoLR (mLR) compared to the human annotation (H). *The MT model may have altered differences in the original Welsh generation.

| Sample | Precision H | Precision mLR | Recall H | Recall mLR | Q |
|---|---|---|---|---|---|
| **Graph:**<br>Mermaid (Train song) \| genre \| Pop rock<br>Mermaid (Train song) \| runtime \| 3.16<br>Mermaid (Train song) \| releaseDate \| 2012-12-27<br>Mermaid (Train song) \| precededBy \| This'll Be My Year<br>Mermaid (Train song) \| writer \| Espen Lind<br>**Evaluated Generation (Eng):**<br>Mermaid is a pop rock song written by Espen Lind. It was released on 27 December 2012 and has a run time of 3.16. | 1.00 | 0.98 | 0.62 | 0.75 | 1st |
| **Graph:**<br>Turkey \| longName \| Republic of Turkey<br>Nurhan Atasoy \| nationality \| Turkish people<br>Nurhan Atasoy \| citizenship \| Turkey<br>Turkey \| language \| Turkish language<br>**Evaluated Generation (Eng):**<br>The Turkish language is spoken in Turkey where the leader is known as the Republic of Turkey. The country is the location of the Ataturk Atasoy which is a citizenship of the Turkish people. | 0.19 | 0.32 | 0.25 | 0.37 | 3rd |
| **Graph:**<br>Ciudad Ayala \| populationMetro \| 1777539<br>**Evaluated Generation (Eng):**<br>1777539 is the population metro in the country. | 0.12 | 0.88 | 0.25 | 0.52 | 5th |

Tab. 5.9: Referenceless metric evaluation on English examples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the accuracy of MonoLR (mLR) compared to the human annotation (H).

### 5.6.3 Retrieval Accuracy

Table 5.10 shows the Accuracy at 1 (A@1) for text/graph retrieval. In all languages, the proposed fine-tuned models outperform the baselines, with the MonoLR model achieving nearly perfect scores in most of them and even surpassing FS in English. While the retrieval corpus is admittedly limited in size (10K possible combinations), the results demonstrate the effect of the proposed approach on multilingual graph-text representation learning: for all languages, they successfully identify the matching text given a graph and vice versa.

| | A@1 | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Breton** | **English** | **Irish** | **Maltese** | **Russian** | **Welsh** | **Xhosa** |
| DQE | 0.53 | 0.95 | 0.32 | 0.48 | 0.05 | 0.39 | 0.38 |
| FS | 0.37 | 0.99 | 0.36 | 0.45 | 0.11 | 0.46 | 0.24 |
| NB | 0.56 | 0.79 | 0.49 | 0.60 | 0.70 | 0.60 | 0.68 |
| MultiFF | 0.92 | **1.00** | 0.84 | **0.96** | 0.96 | 0.95 | **0.99** |
| MultiLR | 0.93 | 0.99 | 0.85 | 0.94 | 0.93 | 0.97 | 0.98 |
| MonoLR | **0.94** | **1.00** | **0.91** | **0.96** | **0.99** | **1.00** | **0.99** |

Tab. 5.10: Referenceless metric Accuracy at 1 (A@1) when using the F1 score to match graphs with their corresponding text on 100 selected examples from the 7L-Auto dataset. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with its closest language (Welsh).*

## 5.7 Conclusion

This chapter studied the third research question by introducing a novel referenceless evaluation metric for RDF-to-Text generation based on Natural Language Inference (NLI). Unlike previous referenceless approaches, which were primarily restricted to English and provided only coarse-grained estimates, the proposed method enables multilingual evaluation. It decomposes semantic faithfulness into interpretable precision, recall, and F1 components.

To support this metric, a new methodology was proposed that enables the creation of annotated training data through automatic graph manipulation, machine translation, and filtering. This approach enables the creation of silver data for both high- and low-resource languages. Empirical results show that these models achieve strong correlation with human judgments of precision and recall, and outperform existing referenceless baselines.

These findings demonstrate that NLI-based models can serve as the foundation for practical, scalable, and interpretable evaluation of multilingual G2T generation, even in settings with scarce resources or no gold-standard references.

# Chapter 6

# Conclusion

As stated in the Chapter 1, despite continued advances, multilingual G2T remains hindered by data scarcity, especially in low-resource settings. This limitation affects the development of multilingual G2T systems in multiple ways. When paired with the data-hungry nature of current training strategies, this lack of sufficient data translates into models with poor generation quality. When paired with the reliance on reference-based and English-centric metrics, it translates into unreliable or outright impossible evaluation settings. This final chapter synthesizes the main findings of this thesis, contextualizes them in terms of the initial research questions, and reflects on their collective impact on the future of multilingual G2T generation and evaluation.

## 6.1 Main Contributions

Below is a summary of the contributions made in the context of each research question:

### RQ1. Can text generation from Resource Description Framework (RDF) graphs be improved in low-resource languages with limited training examples by fully fine-tuning a model with soft prompts enriched with phylogenetic information?

The first research question focused on generating natural language from RDF graphs in low-resource languages. As stated before, while large-scale datasets and powerful models have advanced G2T for high-resource languages, data scarcity remains a critical barrier elsewhere. This thesis hypothesizes that linguistic proximity, languages sharing a family or structural features, can support cross-lingual transfer. Accordingly, it investigates whether structured soft prompts, informed by language relationships, can bolster multilingual models for RDF-to-Text in under-represented settings.

To address this question, the PI-TST (Phylogeny-Inspired Task-Source-Target) soft prompt was introduced. This modular and linguistically structured prompt encodes task, family, genus, and language information for both source and target. These prompts are combined with a multilingual pre-trained model (mT5-large). Training proceeds in three stages: adapting the base model with masked language modeling on monolingual and RDF-specific data, unsupervised pretraining of the prompts using tasks such as prefix or suffix language modeling and deshuffling, and finally fine-tuning on small RDF-to-text datasets.

Experimental results show that PI-TST surpasses both simple full model fine-tuning and strong baselines such as Control Prefixes. The experiments display robust gains on automatic metrics (BLEU, Google BLEU, and LaBSE cosine similarity) and in human evaluations (readability, grammaticality, word order, and semantic adequacy). Notably, the most significant improvements are seen in Breton, a language unseen during the backbone's pretraining, validating the hypothesis that phylogenetic structure supports transfer. Ablation studies reveal that both source and target prompts are necessary for optimal results, and that the method remains data-efficient, requiring as few as 1,000 samples per language to perform competitively.

### RQ2. Can text generation from Abstract Meaning Representation (AMR) graphs be improved using phylogenetic information to guide a model's training process in high- and low-resource languages?

The second research question focused on AMR-to-Text generation and expanded its scope to a broader range of languages, including both high- and low-resource Indo-European languages. Here, the challenge was twofold: limited annotated data in low-resource settings and the risk that multilingual training, while facilitating transfer, may also introduce noise.

To navigate this trade-off, the thesis introduced Hierarchical QLoRA (HQL), a novel curriculum learning strategy that iteratively refines a multilingual model into monolingual ones via parameter-efficient fine-tuning. Building on a 4-bit quantized mT5-large model, LoRA adapters are used to support modular training with low memory and data requirements. Training follows a hierarchy: an all-language model (L0) is refined into a 6-language model (L1), then a bilingual model (L2), and finally a monolingual model (L3). LoRA adapters are reused and extended at each stage, lowering training costs.

Two curricula were explored: one that maximizes language diversity within groups (Distant Language Hierarchy, DLHQL) and another that clusters by linguistic similarity (Phylogenetic Tree Hierarchy, PTHQL). The results show that phylogenetic groupings generally yield the best results, supporting the notion that structured proximity is beneficial for transfer, while also maintaining regularization effects seen with more diverse groups. Both curricula outperform both monolingual and multilingual baselines, as well as a Generate-and-Translate pipeline, especially in low-resource languages such as Asturian and Haitian Creole, but also for high-resource cases.

### RQ3. Can Natural Language Inference (NLI) be used as the base to develop a referenceless multilingual evaluation metric for multiple facets of semantic faithfulness in RDF-to-Text generation across high- and low-resource languages?

The third research question confronted the challenge of evaluation. Most common G2T metrics (BLEU, ChrF++) depend on high-quality reference texts, which are scarce outside of very few languages. At the same time, recent referenceless approaches that aim to address this data scarcity, such as Data-QuestEval and Factspotter, are predominantly English-centric and provide only limited diagnostic value, requiring extensive preprocessing or targeting only omissions.

To handle this issue, the thesis proposed a referenceless metric based on Natural Language Inference (NLI). By taking a multilingual mDeBERTa-v3 model fine-tuned on NLI and further fine-tuning it for regression, the approach estimates semantic precision (text entailed by the graph), recall (graph entailed by the text), and their harmonic mean (F1), between the RDF graph and its generated text. The training data consists of 1.77 million synthetic graph-text

pairs across six languages, constructed by manipulating overlap in WebNLG instances and expanded through machine translation, with extensive filtering for semantic similarity and language identity. Both full fine-tuning and parameter-efficient (LoRA) approaches are evaluated.

Results show a strong correlation between the metric and both human annotations and standard reference-based metrics, even though no gold references are required. The monolingual LoRA variant outperforms previous referenceless metrics across all languages, achieving Spearman correlations up to 0.70 in English and 0.67 in Russian, and performing well in low-resource languages. Critically, the metric decomposes into precision and recall scores, supporting detailed diagnostics for over- and under-generation.

Together, these contributions answer the thesis research questions by demonstrating that phylogenetic information and principled, parameter-efficient training can overcome the key challenges of data scarcity in multilingual G2T. The methods proposed (PI-TST soft prompts, Hierarchical QLoRA, and NLI-based referenceless evaluation) not only advance the technical state of the art but also promote greater linguistic inclusivity and transparency. In doing so, this work provides a scalable blueprint for building and evaluating robust G2T systems across multiple languages.

## 6.2 Limitations

While the findings in this thesis advance multilingual G2T generation and evaluation, it is necessary to acknowledge some limitations.

Much of the empirical progress relies on machine-generated data, given the nonexistence of gold-standard data and the prohibitive cost and time required to create it. While tests were always performed on high-quality data, the use of synthetic data introduces potential systematic biases and limits the generalization of results. Quantifying the impact of these biases remains challenging, yet this highlights the core problem of data scarcity addressed by the thesis.

Language coverage, though broader than in most existing work, remains focused on Indo-European languages. Critical linguistic phenomena, such as rich inflectional morphology, agglutinative structures, or non-Indo-European scripts, remain unexplored. This limitation leaves open the question of whether the proposed techniques transfer robustly to such scenarios.

In terms of analysis and evaluation, interpretability and diagnostic utility are only partially achieved. The thesis introduces an evaluation metric that targets both semantic precision and recall, advancing faithfulness measurement in non-English languages; however, it does not provide localization of errors or actionable ways for systems to address them. When omissions or additions are detected, their specific relation with the input and output remains unclear, which limits applicability for error analysis and user feedback.

Finally, computational and infrastructural constraints have influenced both experimental scope and model choice. Training on larger and more diverse corpora, conducting more systematic cross-lingual analysis, and leveraging recent large language models were all out of reach due to resource limitations, but represent critical future directions as infrastructure evolves.

Together, these limitations shape both the interpretation of current findings and future research aims. They underscore the need for deeper, more balanced, and precise work to advance robust and generalizable multilingual natural language generation (NLG) and evaluation.

## 6.3 Future Research

In light of the limitations above, several research directions emerge.

Reducing biases from synthetic data is crucial. Future work should investigate the improved generation and filtering of synthetic data, as well as the creation of gold-standard resources for more languages, particularly those with low resources.

Expanding both generation and evaluation methods to a broader spectrum of languages, especially non-Indo-European languages, remains essential. This approach would test the robustness and universality of the thesis's methods, especially in languages with complex morphology, agglutinative structures, or underrepresented scripts.

Improving the interpretability and diagnostic value of evaluation metrics remains a significant challenge. Developing token-level or span-level faithfulness attribution and error localization techniques would make evaluation more transparent and actionable for model development and user feedback.

As computational resources become more accessible, future work should revisit large-scale experiments with broader datasets and newer model architectures, including recent large language models. Further exploration of modular and parameter-efficient architectures, such as independently trained source and target LoRA modules or adapters per task or language, may also enhance adaptability and scalability.

In conjunction, these research paths can help create more accessible, reliable, and equitable language technologies, ensuring that the benefits of NLG reach a broader range of language communities and that the scientific study of language achieves a genuinely global scope.

## 6.4 Final Remarks

This thesis aimed to make structured knowledge more accessible and equitable through advances in G2T generation and evaluation, with a particular focus on low-resource languages. By addressing both data scarcity and the limitations of reference-based evaluation, the work aimed to democratize access to information across linguistic boundaries. The proposed methods tackled core challenges in multilingual G2T. In doing so, they confirmed that these approaches enhance both technical quality and inclusivity, supporting reliable and scalable multilingual text generation and evaluation.

Ongoing progress will depend on reducing synthetic data biases, broadening language coverage, improving evaluation interpretability, and further analyzing model architectures. The contributions presented here represent a step towards a more reliable and equitable access to structured knowledge for all language communities.

# Annex A

# Appendices for Chapter 3

## A.1 RDF-to-Text Human Evaluation

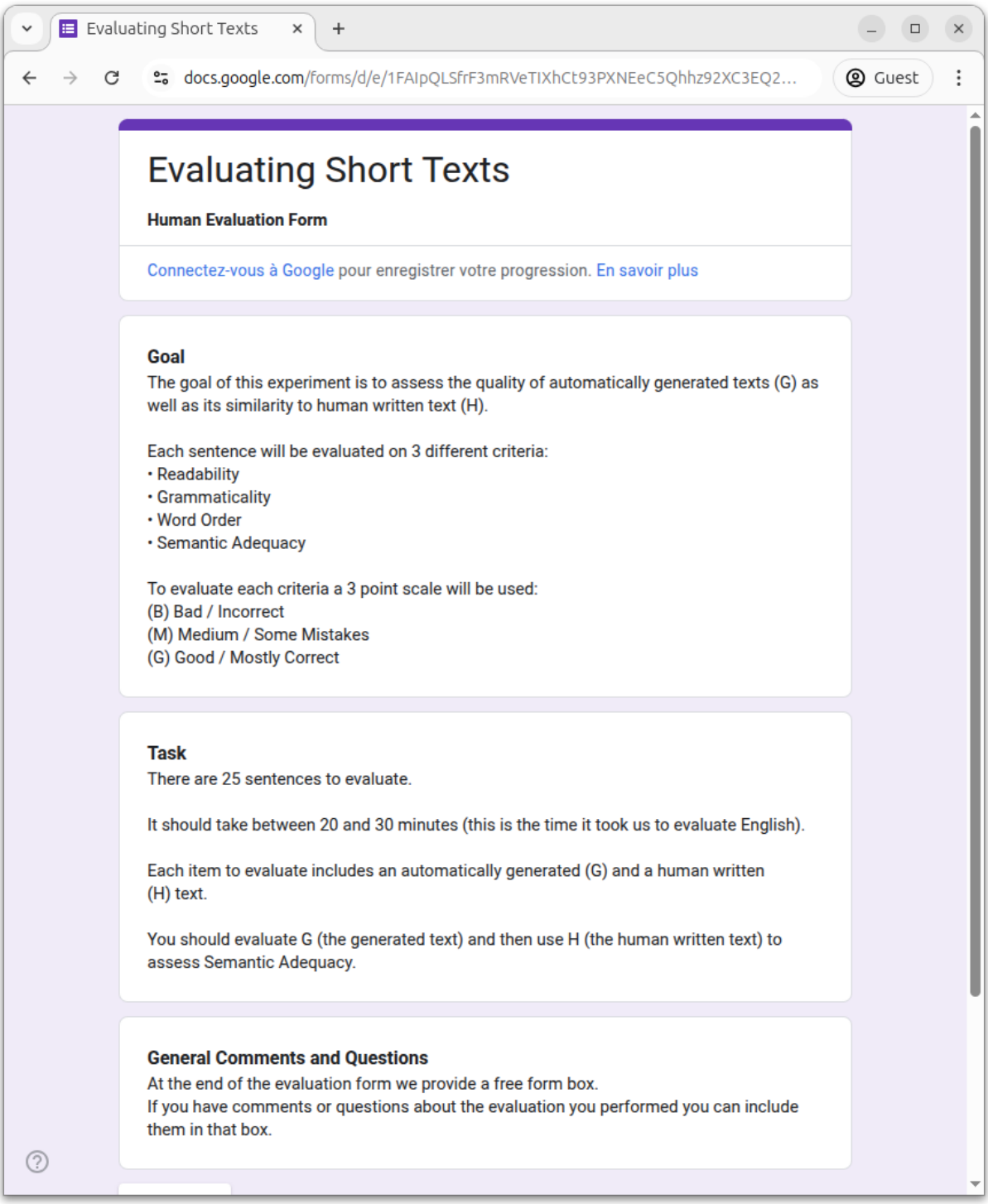



FIG. A.1: Soft prompts human evaluation instructions part 1 of 2

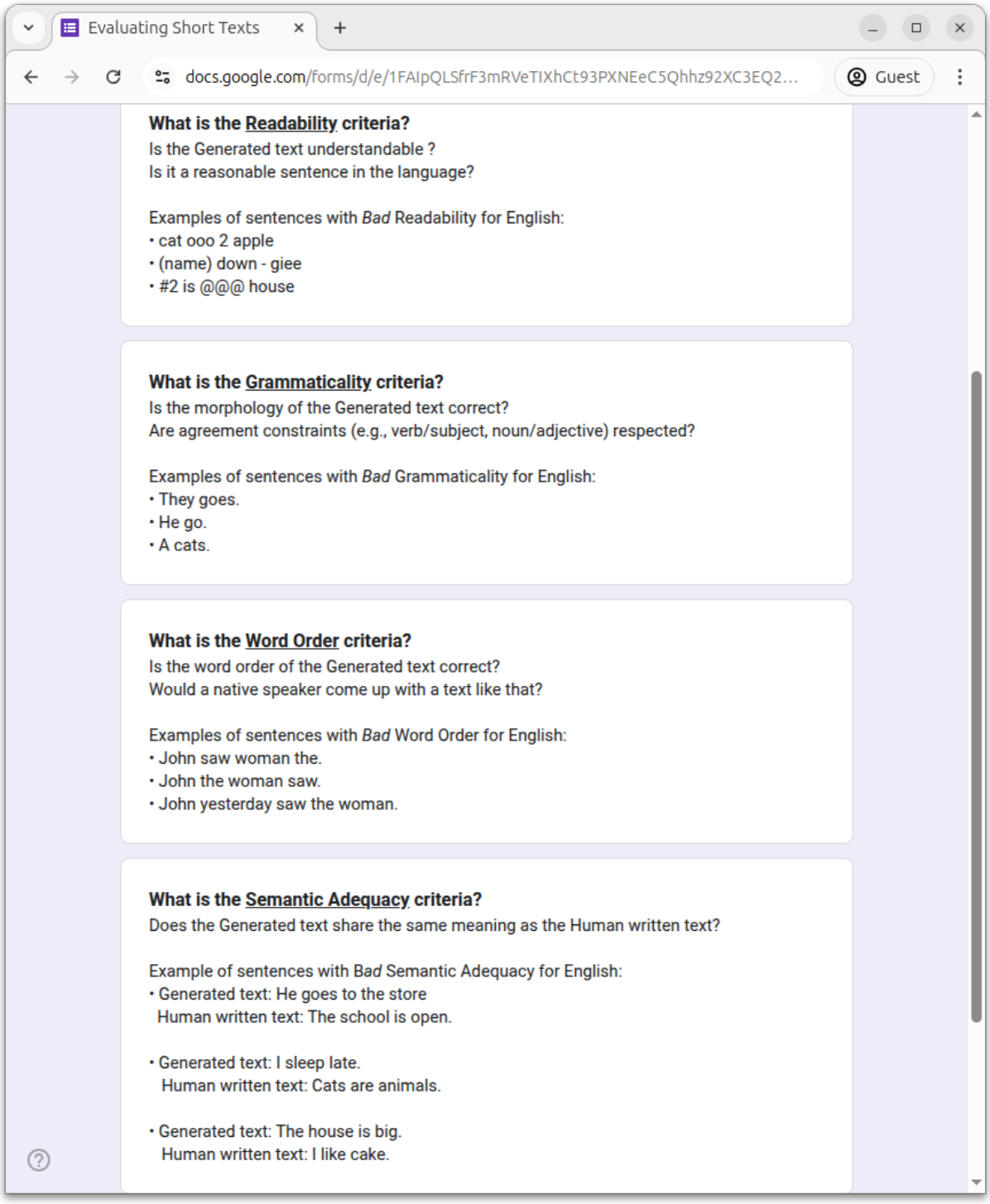



Fig. A.2: Soft prompts human evaluation instructions part 2 of 2

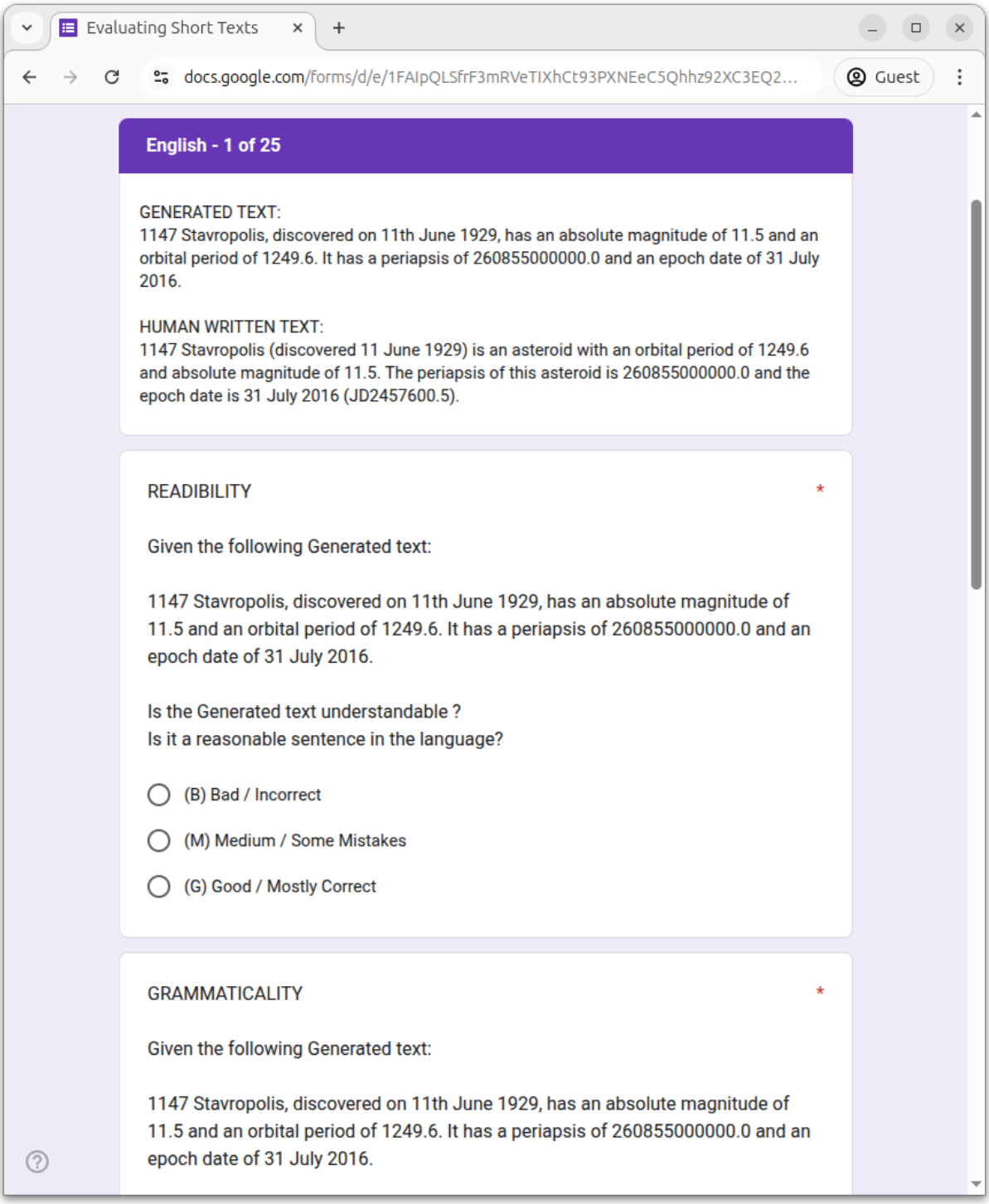



FIG. A.3: Soft prompts human evaluation example question

## A.2 RDF-to-Text Generation Examples

| Sample | R | G | WO | SA | Q |
|---|---|---|---|---|---|
| **Graph:**<br>Liselotte Grschebina \| deathPlace \| Petah Tikva<br>Petah Tikva \| country \| Israel<br>**Generation (Bre):**<br>Liselotte Grschebina zo marvet e Petah Tikva en Israel.<br>**Reference (Bre):**<br>Marvet eo Liselotte Grschebina e Petah Tikva, en Israel.<br>**Generation (Eng MT*):**<br>Liselotte Grschebina died in Petah Tikva, Israel.<br>**Reference (Eng):**<br>Liselotte Grschebina died in Petah Tikva, Israel. | 1.00 | 1.00 | 1.00 | 1.00 | 1st |
| **Graph:**<br>The Hobbit \| literaryGenre \| High fantasy<br>**Generation (Bre):**<br>The Hobbit zo un doare sonerezh kendeuz.<br>**Reference (Bre):**<br>Gourmoliac'h eo rumm lennegezh An Hobbit.<br>**Generation (Eng MT*):**<br>The Hobbit is a fusion of music.<br>**Reference (Eng):**<br>The Hobbit literary genre is high fantasy. | 0.83 | 1.00 | 1.00 | 0.00 | 3rd |
| **Graph:**<br>Ciudad Ayala \| country \| Mexico<br>Mexico \| leaderTitle \| President of the Senate<br>**Generation (Bre):**<br>Ciudad Ayala zo e Mec'hiko lec'h m'emañ ar Prezidant ar Senate ar rener.<br>**Reference (Bre):**<br>E Bro-Vec'hiko emañ Ciudad Ayala ha Prezidant ar Sened eo titl al leader e Bro-Vec'hiko.<br>**Generation (Eng MT*):**<br>Ciudad Ayala is in Mexico where the President of the Senate is the president.<br>**Reference (Eng):**<br>Ciudad Ayala is in the country of Mexico and the President of the Senate is the leader title in Mexico. | 0.66 | 0.33 | 0.50 | 0.50 | 5th |

Tab. A.1: PI-TST Breton generation examples from the human evaluation with the average score across evaluators. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the four human evaluation metrics: Readability (R), grammaticality (G), word order(WO), and semantic adequacy (SA). *The MT model may have altered differences in the original Breton generation.

| Sample | R | G | WO | SA | Q |
|---|---|---|---|---|---|
| **Graph:**<br>Harold French \| birthPlace \| UK<br>**Generation (Gle):**<br>Rugadh Harold French san RA.<br>**Reference (Gle):**<br>Rugadh Harold French sa Ríocht Aontaithe.<br>**Generation (Eng MT*):**<br>Harold French was born in the UK.<br>**Reference (Eng):**<br>Harold French is born in the UK. | 1.00 | 1.00 | 1.00 | 1.00 | 1st |
| **Graph:**<br>Terence Rattigan \| deathYear \| 1977-01-01<br>**Generation (Gle):**<br>Ba ar an 1 Eanáir 1977 a fuair Terence Rattigan bás.<br>**Reference (Gle):**<br>Fuair Terence Rattigan bás i 1977.<br>**Generation (Eng MT*):**<br>Terence Rattigan died on January 1, 1977.<br>**Reference (Eng):**<br>Terence Rattigan died in 1977. | 0.75 | 0.50 | 0.75 | 1.00 | 3rd |
| **Graph:**<br>The Hobbit \| literaryGenre \| High fantasy<br>**Generation (Gle):**<br>Is é an t-alt a bhaineann The Hobbit leis an gceol miotail ard-chreidimh.<br>**Reference (Gle):**<br>The article is about The Hobbit and the high-religion metal music.<br>**Generation (Eng MT*):**<br>Liselotte Grschebina died in Petah Tikva, Israel.<br>**Reference (Eng):**<br>The Hobbit literary genre is high fantasy. | 0.00 | 0.50 | 0.50 | 0.25 | 5th |

Tab. A.2: PI-TST Irish generation examples from the human evaluation with the average score across evaluators. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the four human evaluation metrics: Readability (R), grammaticality (G), word order(WO), and semantic adequacy (SA). *The MT model may have altered differences in the original Irish generation.

# Annex B

# Appendices for Chapter 4

## B.1 AMR-to-Text Generation Examples

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Ltz):**<br>Mee et gëtt nach ëmmer vill Saachen, déi wéi Dinosaurier ausgesinn.<br>**Reference (Ltz):**<br>Awer et gi vill Saache bei Vullen, déi ëmmer nach ewéi en Dinosaurier ausgesinn.<br>**Generation (Eng MT*):**<br>But there are still many things that look like dinosaurs.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.13 | 0.52 | 0.69 | 1st |
| **Graph:**<br>(r / religion<br>:name (n / name<br>:op1 "Orthodox"<br>:op2 "Christian")<br>:ARG1-of (m / major-02)<br>:location (c / country<br>:name (n2 / name<br>:op1 "Moldova")))<br>**Generation (Ltz):**<br>D'Orthodoxesch Chrëschtentum ass eng grouss Relioun an Moldawien.<br>**Reference (Ltz):**<br>D'Haaptrelioun a Moldawien ass Chrëschtlech Orthodox.<br>**Generation (Eng MT*):**<br>Orthodox Christianity is a major religion in Moldova.<br>**Reference (Eng):**<br>The major religion in Moldova is Orthodox Christian. | 0.06 | 0.43 | 0.69 | 3rd |
| **Graph:**<br>(p / possible-01<br>:ARG1 (n / need-01<br>:ARG0 (y2 / you)<br>:ARG1 (y / yacht)<br>:location (a / and<br>:op1 (i / island)<br>:op2 (l / lake))<br>:polarity -))<br>**Generation (Ltz):**<br>Dir kënnt net e Yacht op Inselen a Länner brauchen.<br>**Reference (Ltz):**<br>An den Archipeler a Séie brauch een net zwéngend eng Yacht.<br>**Generation (Eng MT*):**<br>You may not need a yacht on islands and countries.<br>**Reference (Eng):**<br>In the archipelagos and lakes you do not necessarily need a yacht. | 0.06 | 0.24 | 0.41 | 5th |

Tab. B.1: PTHQL Luxembourgish generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Luxembourgish generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(r / religion<br>:name (n / name<br>:op1 "Orthodox"<br>:op2 "Christian")<br>:ARG1-of (m / major-02)<br>:location (c / country<br>:name (n2 / name<br>:op1 "Moldova")))<br>**Generation (Deu):**<br>Die größte Religion in Moldawien ist die Orthodoxe.<br>**Reference (Deu):**<br>Die wichtigste Religion in Moldawien ist die christlich-orthodoxe.<br>**Generation (Eng MT*):**<br>The largest religion in Moldova is Orthodox.<br>**Reference (Eng):**<br>The major religion in Moldova is Orthodox Christian. | 0.48 | 0.61 | 0.82 | 1st |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Deu):**<br>Die Aufnahmegebühr für dieses Bildungsprogramm ist normalerweise eine Studiengebühr.<br>**Reference (Deu):**<br>In der Regel wird für die Anmeldung zu diesen Bildungsprogrammen eine Studiengebühr erhoben.<br>**Generation (Eng MT*):**<br>The admission fee for this educational program is usually a tuition fee.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.07 | 0.42 | 0.83 | 3rd |
| **Graph:**<br>(p / possible-01<br>:ARG1 (n / need-01<br>:ARG0 (y2 / you)<br>:ARG1 (y / yacht)<br>:location (a / and<br>:op1 (i / island)<br>:op2 (l / lake))<br>:polarity -))<br>**Generation (Deu):**<br>Sie brauchen keine Yachten auf Inseln und Flüssen.<br>**Reference (Deu):**<br>Für die Archipele und Seen brauchen Sie nicht unbedingt eine Yacht.<br>**Generation (Eng MT*):**<br>You don't need yachts on islands and rivers.<br>**Reference (Eng):**<br>In the archipelagos and lakes you do not necessarily need a yacht. | 0.05 | 0.28 | 0.67 | 5th |

Tab. B.2: PTHQL German generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original German generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(a / and<br>:op1 (d2 / display-01<br>:ARG1 (s / scene)<br>:ARG2 (p / pyramid))<br>:op2 (l / light-up-08<br>:ARG1 (p2 / pyramid<br>:ARG1-of (d / differ-02))))<br>**Generation (Lim):**<br>De scéne weurt in piramides oetgeveurd en versjèllende piramides weure opgelichte.<br>**Reference (Lim):**<br>De scènes weure op de piramides getoend en de aander piramides weurde verleech.<br>**Generation (Eng MT*):**<br>The scene is set in pyramids and several pyramids are lit up.<br>**Reference (Eng):**<br>The scenes are displayed on the pyramids and the different pyramids are lit up. | 0.04 | 0.44 | 0.41 | 1st |
| **Graph:**<br>(w / worth-02<br>:ARG1 (s / stroll-01<br>:ARG1 (a / about<br>:op1 (v / village<br>:ARG0-of (i / intrigue-01))))<br>:ARG2 (t / temporal-quantity<br>:quant 0.5<br>:unit (h / hour)))<br>**Generation (Lim):**<br>'n Half uur gaon door 't intrigerende dörp.<br>**Reference (Lim):**<br>'t Loent ziech de meujte um e haaf oor door 't intrigerende dörp te wandele.<br>**Generation (Eng MT*):**<br>Half an hour walk through the intriguing village.<br>**Reference (Eng):**<br>It's worth half an hour to stroll about the intriguing village. | 0.05 | 0.29 | 0.34 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Lim):**<br>'n Typische studiefooi is veur 't insjrieve in dit oonderwiesprogramma.<br>**Reference (Lim):**<br>Miestal moot me collegegeld betaole um aongenome te weure in dees studieprogramma's.<br>**Generation (Eng MT*):**<br>A typical study sheet is for enrollment in this educational program.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.04 | 0.21 | 0.25 | 5th |

TAB. B.3: PTHQL Limburgish generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Limburgish generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(b / build-01<br>:ARG1 (r / railway)<br>:location (c / country<br>:name (n / name<br>:op1 "England"))<br>:time (d / date-entity<br>:century 16<br>:mod (e / early)))<br>**Generation (Nld):**<br>In het vroege 16e eeuw werden spoorwegen in Engeland gebouwd.<br>**Reference (Nld):**<br>Al in de 16e eeuw werden er in Engeland paardenspoorwegen gebouwd.<br>**Generation (Eng MT*):**<br>Railways were built in England in the early 16th century.<br>**Reference (Eng):**<br>Wagonways were built in England as early as the 16th Century. | 0.18 | 0.62 | 0.63 | 1st |
| **Graph:**<br>(r / religion<br>:name (n / name<br>:op1 "Orthodox"<br>:op2 "Christian")<br>:ARG1-of (m / major-02)<br>:location (c / country<br>:name (n2 / name<br>:op1 "Moldova")))<br>**Generation (Nld):**<br>De belangrijkste religie in Moldavië is de Orthodoxe Christendom.<br>**Reference (Nld):**<br>Orthodox-christelijk is de voornaamste godsdienst in Moldavië.<br>**Generation (Eng MT*):**<br>The main religion in Moldova is Orthodox Christianity.<br>**Reference (Eng):**<br>The major religion in Moldova is Orthodox Christian. | 0.12 | 0.42 | 0.87 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Nld):**<br>Een typische studiekostenvergoeding is voor het inschrijven van dit onderwijsprogramma.<br>**Reference (Nld):**<br>Normaliter moet er voor de inschrijving voor deze educatieve programma's collegegeld worden betaald.<br>**Generation (Eng MT*):**<br>A typical tuition fee reimbursement is for enrolling in this educational program.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.03 | 0.23 | 0.53 | 5th |

Tab. B.4: PTHQL Dutch generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Dutch generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(w / way :ARG1-of (i / important-01)<br>:domain (t / this)<br>:manner-of (d / distinguish-01<br>:ARG1 (a / and<br>:op1 (v / verb<br>:mod (s / some))<br>:op2 (o / object<br>:mod (s2 / some)))))<br>**Generation (Ast):**<br>Esta ye una forma importante de distinguir dellos verbos y oxetos.<br>**Reference (Ast):**<br>Ye un mou importante d'estremar ente dellos verbos y oxetos.<br>**Generation (Eng MT*):**<br>This is an important way to distinguish between verbs and objects.<br>**Reference (Eng):**<br>This is an important way to distinguish between some verbs and objects. | 0.33 | 0.57 | 0.67 | 1st |
| **Graph:**<br>(p / possible-01<br>:ARG1 (t / theme<br>:ARG1-of (g / good-02)<br>:ARG2-of (b / base-02<br>:ARG1 (h / holiday-01)<br>:location (a / around))<br>:domain (w / waterway<br>:ARG1-of (l / land-01))))<br>**Generation (Ast):**<br>La ruta d'alcuerdu puede ser un bon tema pa unas vacaciones.<br>**Reference (Ast):**<br>Les canales d'interior son un bon tema de viaxe.<br>**Generation (Eng MT*):**<br>The route of agreement can be a good theme for a vacation.<br>**Reference (Eng):**<br>Inland waterways can be a good theme to base a holiday around. | 0.14 | 0.31 | 0.37 | 3rd |
| **Graph:**<br>(p / possible-01<br>:ARG1 (n / need-01<br>:ARG0 (y2 / you)<br>:ARG1 (y / yacht)<br>:location (a / and<br>:op1 (i / island)<br>:op2 (l / lake))<br>:polarity -))<br>**Generation (Ast):**<br>Nun hai necesidá d'un yate na islla y nel llagu.<br>**Reference (Ast):**<br>Nos archipiélagos y nos llagos nun ye necesario tener un yate.<br>**Generation (Eng MT*):**<br>There is no need for a yacht on the island and the lagoon.<br>**Reference (Eng):**<br>In the archipelagos and lakes you do not necessarily need a yacht. | 0.04 | 0.22 | 0.36 | 5th |

Tab. B.5: PTHQL Asturian generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The used MT model does not support Asturian, Spa-to-Eng translation was used instead. The MT model may have altered differences in the original Asturian generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Spa):**<br>La matrícula es típicamente una cuota de matrícula para inscribirse en este programa educativo.<br>**Reference (Spa):**<br>Normalmente, se cobrará una tarifa de matrícula para inscribirse en estos programas educativos.<br>**Generation (Eng MT*):**<br>Tuition is typically a registration fee to enroll in this educational program.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.27 | 0.54 | 0.66 | 1st |
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Spa):**<br>Pero hay muchas cosas que todavía parecen dinosaurios en los aves.<br>**Reference (Spa):**<br>Sin embargo, hay muchas características en las aves que todavía las asemejan a los dinosaurios.<br>**Generation (Eng MT*):**<br>But there are many things that still seem dinosaur-like about birds.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.08 | 0.41 | 0.65 | 3rd |
| **Graph:**<br>(t / think-01<br>:ARG0 (y / you)<br>:ARG1 (r2 / route<br>:mod (s / ski-01))<br>:ARG2 (r3 / route<br>:purpose (h / hike-01)<br>:ARG1-of (r / resemble-01<br>:ARG2 r2))<br>:mode imperative)<br>**Generation (Spa):**<br>Piensa en las rutas de esquia como rutas de caminata.<br>**Reference (Spa):**<br>Imagínese la pista de esquí como una ruta de senderismo.<br>**Generation (Eng MT*):**<br>Think of ski trails as hiking trails.<br>**Reference (Eng):**<br>Think of the skiing route as of a similar hiking route. | 0.05 | 0.24 | 0.62 | 5th |

Tab. B.6: PTHQL Spanish generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Spanish generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Hat):**<br>Men, gen anpil bagay sou zwazo ki toujou sanble ak dinozò.<br>**Reference (Hat):**<br>Men gen anpil bagay sou zwazo yo ki sanble ankò ak yon dinozò.<br>**Generation (Eng MT*):**<br>But there are many things about birds that still resemble dinosaurs.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.34 | 0.62 | 0.83 | 1st |
| **Graph:**<br>(p / possible-01<br>:ARG1 (n / need-01<br>:ARG0 (y2 / you)<br>:ARG1 (y / yacht)<br>:location (a / and<br>:op1 (i / island)<br>:op2 (l / lake))<br>:polarity -))<br>**Generation (Hat):**<br>Ou pa ka bezwen yon yacht sou zile yo ak lak yo.<br>**Reference (Hat):**<br>Nan achipèl ak lak yo, li pa nesesè pou gen yon yacht.<br>**Generation (Eng MT*):**<br>You may not need a yacht on the islands and lakes.<br>**Reference (Eng):**<br>In the archipelagos and lakes you do not necessarily need a yacht. | 0.14 | 0.36 | 0.54 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Hat):**<br>Yon frè ekolaj tipik pou enskri nan pwogram edikasyon sa a.<br>**Reference (Hat):**<br>An règ jeneral, enskripsyon pou pwogram ansèyman sa yo mennen ak depans pou eskolarite a.<br>**Generation (Eng MT*):**<br>A typical tuition fee to enroll in this educational program.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.06 | 0.24 | 0.36 | 5th |

Tab. B.7: PTHQL Haitian Creole generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Haitian Creole generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(c / contrast-01<br>:ARG2 (t / thing<br>:quant (l2 / lot)<br>:ARG0-of (l / look-02<br>:ARG1 (d / dinosaur)<br>:mod (s / still))<br>:topic (b / bird)))<br>**Generation (Fra):**<br>Mais il y a beaucoup de choses qui ressemblent à des dinosaures quand il s'agit d'oiseaux.<br>**Reference (Fra):**<br>Mais il y a beaucoup de choses sur les oiseaux qui ressemblent encore à un dinosaure.<br>**Generation (Eng MT*):**<br>But there are a lot of things that look like dinosaurs when it comes to birds.<br>**Reference (Eng):**<br>But there are a lot of things about birds that still look like a dinosaur. | 0.40 | 0.64 | 0.70 | 1st |
| **Graph:**<br>(p / possible-01<br>:ARG1 (t / theme<br>:ARG1-of (g / good-02)<br>:ARG2-of (b / base-02<br>:ARG1 (h / holiday-01)<br>:location (a / around))<br>:domain (w / waterway<br>:ARG1-of (l / land-01))))<br>**Generation (Fra):**<br>Les routes d'atterrissage peuvent être un bon thème pour les vacances.<br>**Reference (Fra):**<br>Les voies navigables intérieures peuvent être un excellent thème pour des vacances.<br>**Generation (Eng MT*):**<br>Landing routes can be a good theme for a vacation.<br>**Reference (Eng):**<br>Inland waterways can be a good theme to base a holiday around. | 0.18 | 0.46 | 0.50 | 3rd |
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Fra):**<br>Il y a généralement une frais de scolarité pour s'inscrire dans ce programme d'éducation.<br>**Reference (Fra):**<br>En règle générale, il faut payer des frais d'inscription pour suivre ces programmes éducatifs.<br>**Generation (Eng MT*):**<br>There is usually a tuition fee to enroll in this education program.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.05 | 0.24 | 0.62 | 5th |

Tab. B.8: PTHQL French generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original French generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(a / and<br>:op1 (d2 / display-01<br>:ARG1 (s / scene)<br>:ARG2 (p / pyramid))<br>:op2 (l / light-up-08<br>:ARG1 (p2 / pyramid<br>:ARG1-of (d / differ-02))))<br>**Generation (Scn):**<br>Li piramidi sunnu 'n vitrina e li diversi piramidi sunnu illuminati.<br>**Reference (Scn):**<br>Li sceni sunnu prujittati ncapu ê piràmidi e li vari piràmidi sunnu illuminati.<br>**Generation (Eng MT*):**<br>The pyramids are on display and the different pyramids are illuminated.<br>**Reference (Eng):**<br>The scenes are displayed on the pyramids and the different pyramids are lit up. | 0.14 | 0.44 | 0.33 | 1st |
| **Graph:**<br>(s2 / summarize-01<br>:ARG1 (s / situation<br>:location (c / country<br>:quant 1)<br>:mod (p / politics))<br>:ARG2 (a / advise-01)<br>:duration (b / brief)<br>:mod (m / mere))<br>**Generation (Scn):**<br>Un breve riassunto della situazione politica in un paese.<br>**Reference (Scn):**<br>Gli avvisi costituiscono semplicemente un breve riepilogo della situazione politica di un Paese.<br>**Generation (Eng MT*):**<br>A brief summary of the political situation in a country.<br>**Reference (Eng):**<br>Advisories are merely a brief summary of the political situation in one country. | 0.12 | 0.40 | 0.64 | 3rd |
| **Graph:**<br>(w / worth-02<br>:ARG1 (s / stroll-01<br>:ARG1 (a / about<br>:op1 (v / village<br>:ARG0-of (i / intrigue-01))))<br>:ARG2 (t / temporal-quantity<br>:quant 0.5<br>:unit (h / hour)))<br>**Generation (Scn):**<br>Un viaggiu ntô paisi intriganti vali na mitati di ura.<br>**Reference (Scn):**<br>Vali la pena fàrisi na passijata di menz'ura ntô villaggiu.<br>**Generation (Eng MT*):**<br>A trip through the intriguing village is worth half an hour.<br>**Reference (Eng):**<br>It's worth half an hour to stroll about the intriguing village. | 0.05 | 0.22 | 0.10 | 5th |

Tab. B.9: PTHQL Sicilian generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Sicilian generation. **The graphs were automatically parsed and may contain errors.

| Sample | B | C | B20 | Q |
|---|---|---|---|---|
| **Graph:**<br>(f / fee<br>:purpose (e2 / enroll-01<br>:ARG2 (p / program<br>:mod (e / educate-01)<br>:mod (t / this)))<br>:ARG1-of (t3 / typical-02)<br>:mod (t2 / tuition))<br>**Generation (Ita):**<br>Una tassa di iscrizione a questi programmi educativi è di solito una tassa di iscrizione.<br>**Reference (Ita):**<br>Di solito l'iscrizione a questi programmi educativi richiede il pagamento di una tassa.<br>**Generation (Eng MT*):**<br>An enrollment fee for these educational programs is usually a tuition fee.<br>**Reference (Eng):**<br>Typically there will be a tuition fee to enroll in these educational programs. | 0.20 | 0.58 | 0.61 | 1st |
| **Graph:**<br>(c / check-01<br>:ARG0 (y / you)<br>:ARG1 (t2 / thing<br>:ARG2-of (l / label-01))<br>:ARG2 (i / instruct-01<br>:ARG1 (p / poison-01<br>:ARG1-of (s2 / specific-02)<br>:mod (t / that))<br>:ARG1-of (s / specific-02)<br>:ARG2 (a / aid-01<br>:mod (f / first))))<br>**Generation (Ita):**<br>Check l'etichetta pi l'istruzioni specifici di l'aiutu primu pi stu poison specificu.<br>**Reference (Ita):**<br>Cuntrolla l'etichetta pi l'istruzzioni pû primu succursu pi chiḍḍu velenu spicìficu.<br>**Generation (Eng MT*):**<br>Check the label for specific first aid instructions for this specific poison.<br>**Reference (Eng):**<br>Check the label for specific first aid instructions for that specific poison. | 0.04 | 0.34 | 0.23 | 3rd |
| **Graph:**<br>(p / possible-01<br>:ARG1 (n / need-01<br>:ARG0 (y2 / you)<br>:ARG1 (y / yacht)<br>:location (a / and<br>:op1 (i / island)<br>:op2 (l / lake))<br>:polarity -))<br>**Generation (Ita):**<br>Non si potrebbe avere bisogno di un yacht su un'isola e un lago.<br>**Reference (Ita):**<br>Non serve per forza uno yacht per gli arcipelaghi e i laghi.<br>**Generation (Eng MT*):**<br>You wouldn't need a yacht on an island and a lake.<br>**Reference (Eng):**<br>In the archipelagos and lakes you do not necessarily need a yacht. | 0.04 | 0.19 | 0.54 | 5th |

Tab. B.10: PTHQL Italian generation examples and their score from automatic metrics. The samples were selected from the 1st (best), 3rd (medium), and 5th (worst) quintile (Q) based on the average of the three automatic metrics: BLEU (B), ChrF++ (C), and BLEURT-20 (B20). *The MT model may have altered differences in the original Italian generation. **The graphs were automatically parsed and may contain errors.

# Annex C

# Appendices for Chapter 5

## C.1 Referenceless metric synthetic dataset creation example

Starting with the following dataset of aligned (graph $g_i$, text $t_i$) pairs:

| ID | Graph | Text |
|---|---|---|
| 1 | Alice \| occupation \| Writer | Alice is a writer. |
| 2 | Alice \| occupation \| Writer<br>Alice \| country \| USA | Alice is an American writer. |
| 3 | Alice \| country \| USA<br>Bob \| country \| USA | Alice and Bob are Americans. |

It is possible to assign all of them precision and recall values of 1, since they are aligned, and turn them into quadruples.

| ID | Graph | Text | P | R |
|---|---|---|---|---|
| 1 | Alice \| occupation \| Writer | Alice is a writer. | 1.00 | 1.00 |
| 2 | Alice \| occupation \| Writer<br>Alice \| country \| USA | Alice is an American writer. | 1.00 | 1.00 |
| 3 | Alice \| country \| USA<br>Bob \| country \| USA | Alice and Bob are Americans. | 1.00 | 1.00 |

To create new quadruples, start by pairing texts with subgraphs or supergraphs of their original graph. For example, $g_1$ is a subgraph of $g_2$ (and therefore $g_2$ is a supergraph of $g_1$). Pairing a text with a supergraph will produce a quadruple where the text is missing information (omission), leading to a lower recall. Pairing a text with a subgraph will produce a quadruple where the text has extra information (addition/hallucination), leading to a lower precision:

| ID | Graph | Text | P | R |
|---|---|---|---|---|
| 1 | Alice \| occupation \| Writer | Alice is a writer. | 1.00 | 1.00 |
| 2 | Alice \| occupation \| Writer<br>Alice \| country \| USA | Alice is an American writer. | 1.00 | 1.00 |
| 3 | Alice \| country \| USA<br>Bob \| country \| USA | Alice and Bob are Americans. | 1.00 | 1.00 |
| 4 | Alice \| occupation \| Writer | Alice is an **American** writer. | 0.50 | 1.00 |
| 5 | Alice \| occupation \| Writer<br>**Alice \| country \| USA** | Alice is a writer. | 1.00 | 0.50 |

It is also possible to pair a text with partially overlapping graphs. For example, $g_2$ and $g_3$ overlap in one triple, matching their texts and graphs creates new quadruples where both recall and precision can be affected (there are both omissions and additions/hallucinations):

| ID | Graph | Text | P | R |
|---|---|---|---|---|
| 1 | Alice \| occupation \| Writer | Alice is a writer. | 1.00 | 1.00 |
| 2 | Alice \| occupation \| Writer<br>Alice \| country \| USA | Alice is an American writer. | 1.00 | 1.00 |
| 3 | Alice \| country \| USA<br>Bob \| country \| USA | Alice and Bob are Americans. | 1.00 | 1.00 |
| 4 | Alice \| occupation \| Writer | Alice is an **American** writer. | 0.50 | 1.00 |
| 5 | Alice \| occupation \| Writer<br>**Alice \| country \| USA** | Alice is a writer. | 1.00 | 0.50 |
| 6 | Alice \| country \| USA<br>**Bob \| country \| USA** | Alice is an American **writer.** | 0.50 | 0.50 |
| 7 | **Alice \| occupation \| Writer**<br>Alice \| country \| USA | Alice **and Bob** are Americans. | 0.50 | 0.50 |

Finally, new quadruples can be created by making synthetic graphs, either by corrupting original triples or by adding new ones:

| ID | Graph | Text | P | R |
|---|---|---|---|---|
| 1 | Alice \| occupation \| Writer | Alice is a writer. | 1.00 | 1.00 |
| 2 | Alice \| occupation \| Writer<br>Alice \| country \| USA | Alice is an American writer. | 1.00 | 1.00 |
| 3 | Alice \| country \| USA<br>Bob \| country \| USA | Alice and Bob are Americans. | 1.00 | 1.00 |
| 4 | Alice \| occupation \| Writer | Alice is an **American** writer. | 0.50 | 1.00 |
| 5 | Alice \| occupation \| Writer<br>**Alice \| country \| USA** | Alice is a writer. | 1.00 | 0.50 |
| 6 | Alice \| country \| USA<br>**Bob \| country \| USA** | Alice is an American **writer.** | 0.50 | 0.50 |
| 7 | **Alice \| occupation \| Writer**<br>Alice \| country \| USA | Alice **and Bob** are Americans. | 0.50 | 0.50 |
| 8 | Alice \| occupation \| Writer<br>Alice \| country \| **Mexico** | Alice is an **American** writer. | 0.50 | 0.50 |
| 9 | Alice \| occupation \| Writer<br>Alice \| country \| USA<br>**Alice \| birthDate \| 2000-01-01** | Alice is an American writer. | 1.00 | 0.66 |

## C.2 4L-RP-Human Annotation

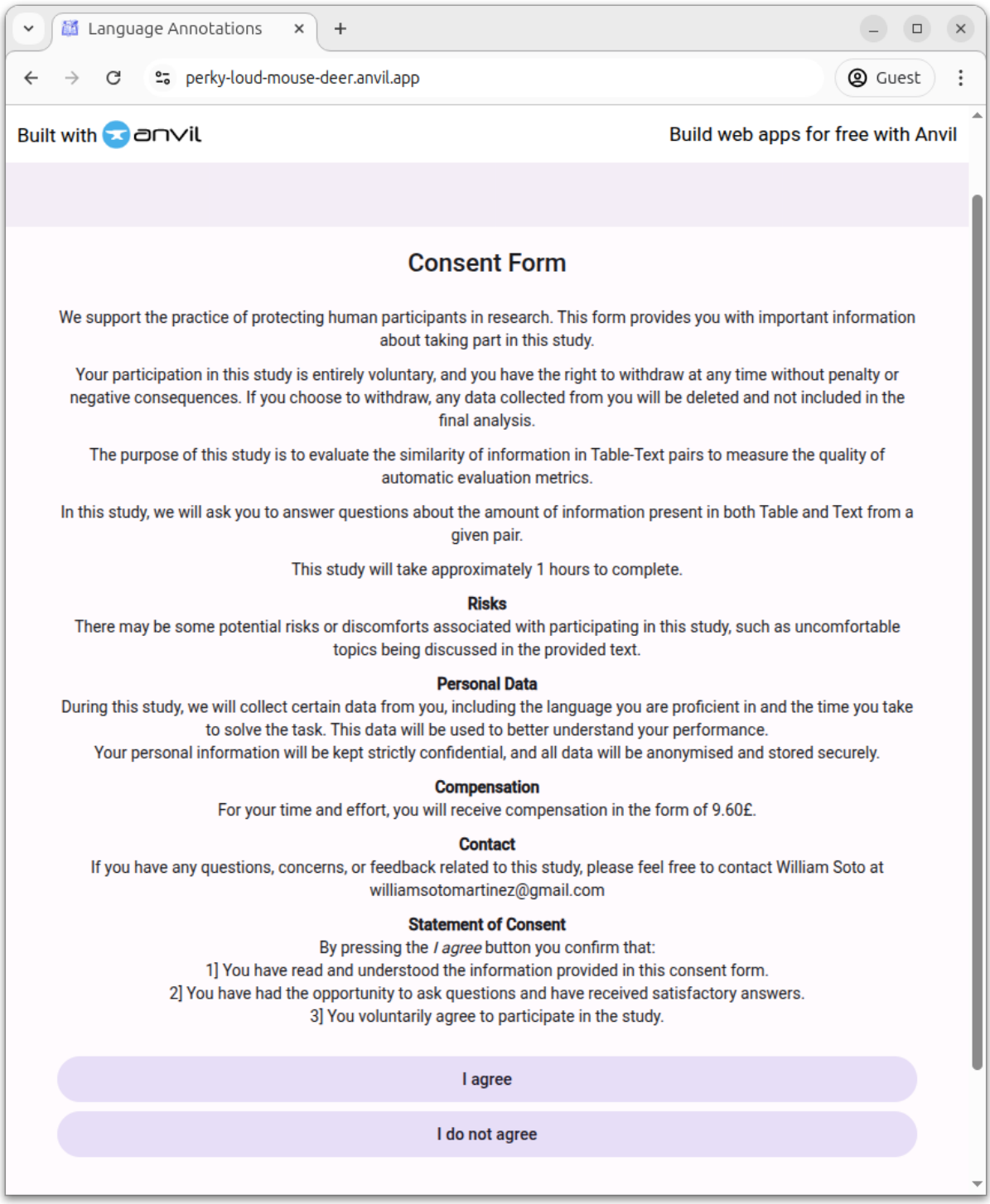



Fig. C.1: 4L-RP-Human annotation consent form part 1 of 2

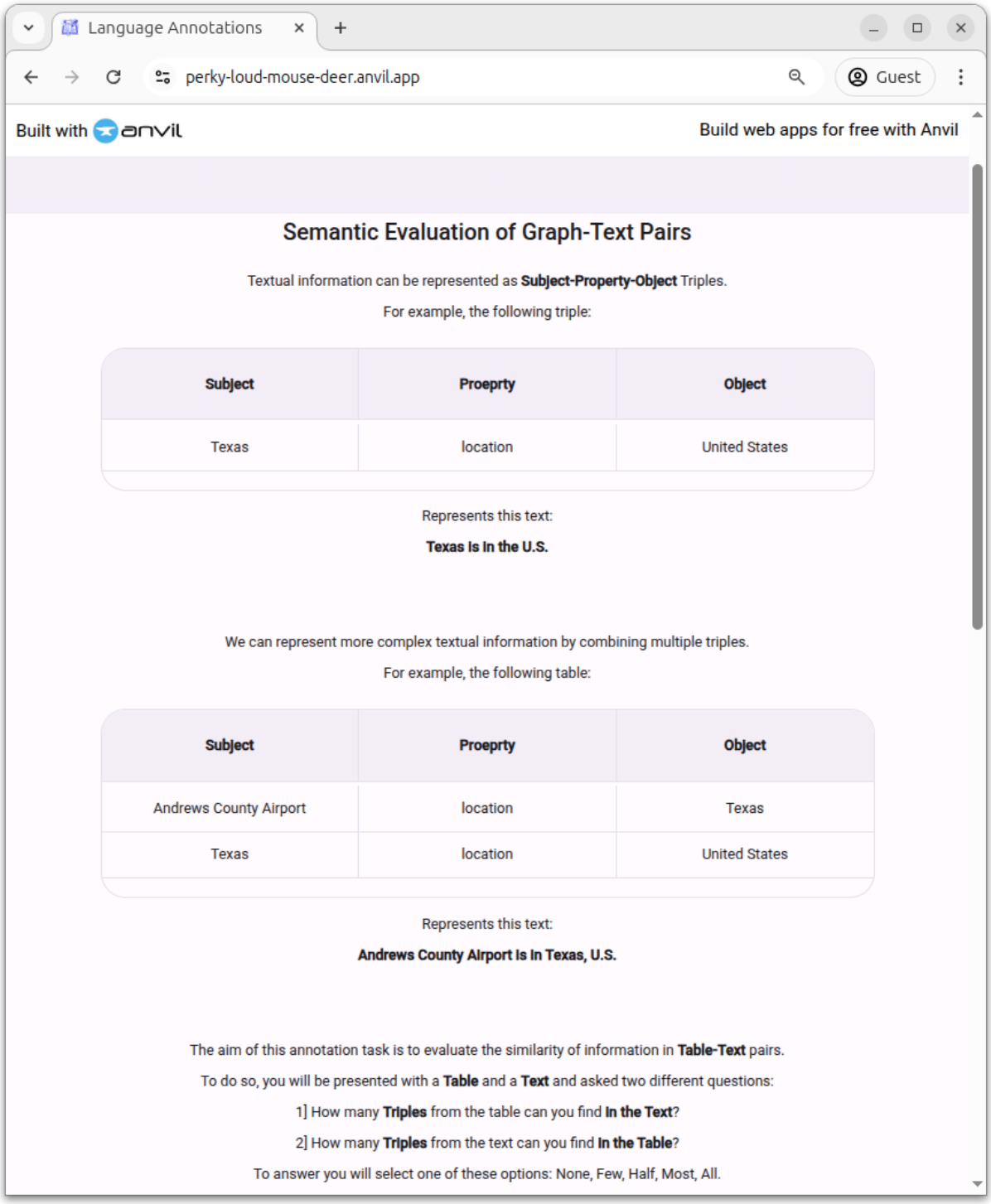



Fig. C.2: 4L-RP-Human annotation instructions part 1 of 2

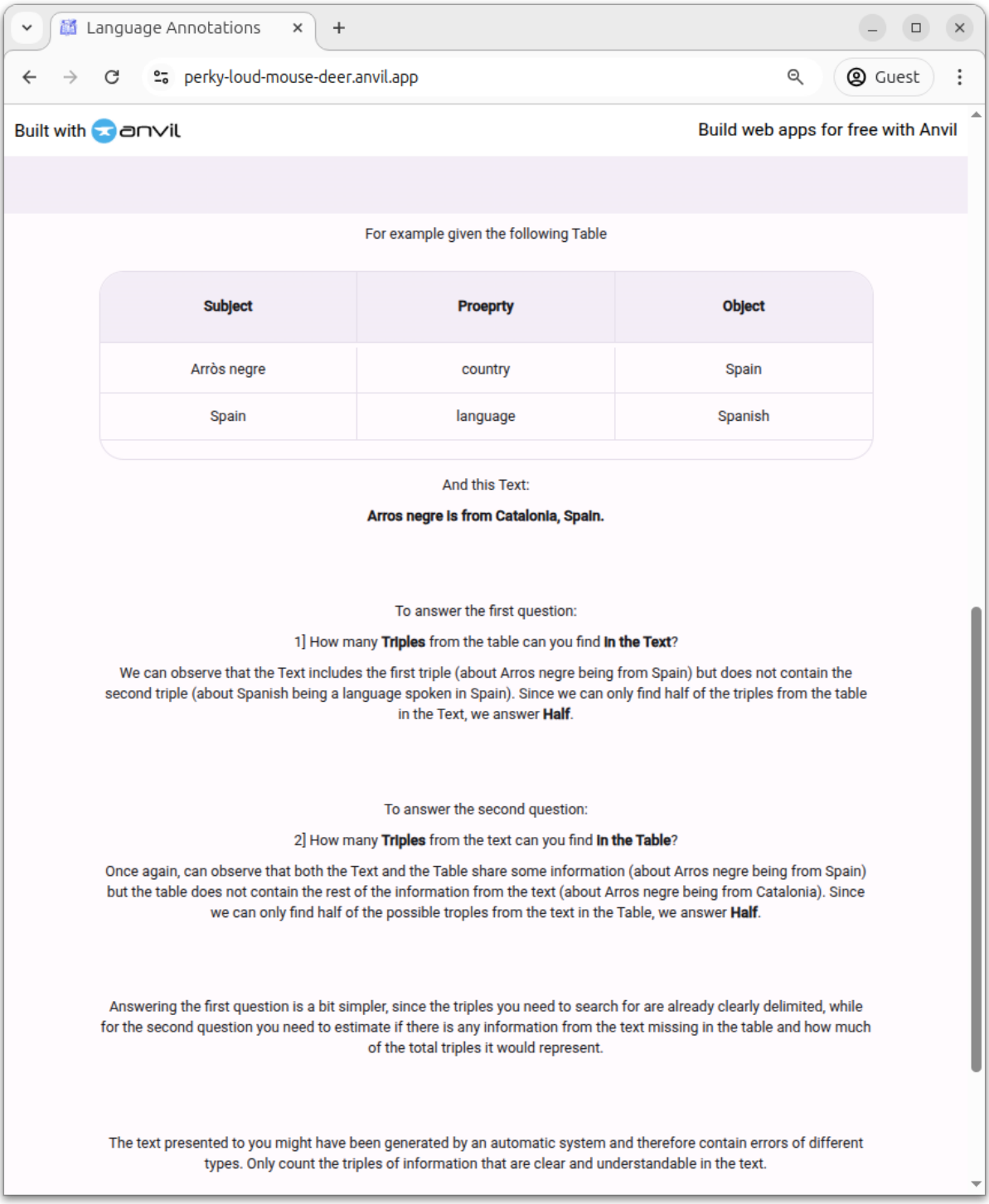



Fig. C.3: 4L-RP-Human annotation instructions part 2 of 2

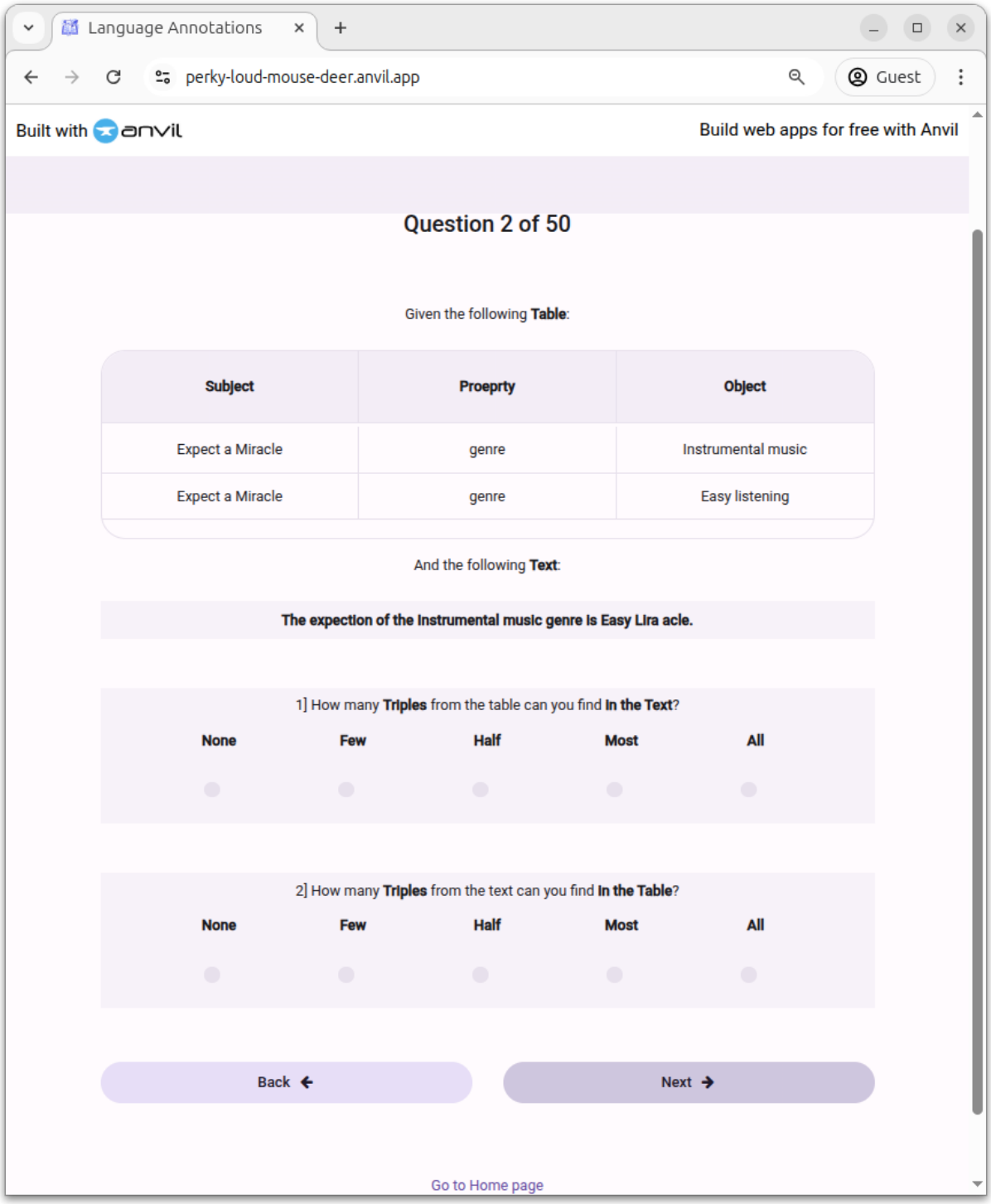



Fig. C.4: 4L-RP-Human annotation example question

## C.3 Referenceless Metric Evaluation Examples

| Sample | Precision H | Precision mLR | Recall H | Recall mLR | Q |
|---|---|---|---|---|---|
| **Graph:**<br>McVeagh of the South Seas \| director \| Cyril Bruce<br>McVeagh of the South Seas \| writer \| Harry Carey (actor born 1878)<br>**Evaluated Generation (Mlt):**<br>McVeagh tal-Baħar tan-Nofsinhar kien miktub minn Harry Carey (attur imwieled 1878) u dirett minn Cyril Bruce.<br>**Evaluated Generation (Eng MT*):**<br>McVeagh of the Southern Seaboard was written by Harry Carey (actor born 1878) and directed by Cyril Bruce. | 0.92 | 0.96 | 0.83 | 0.88 | 1st |
| **Graph:**<br>United States \| leaderTitle \| Vice President<br>Darinka Dentcheva \| residence \| United States<br>**Evaluated Generation (Mlt):**<br>Il-Viċi President huwa l-mexxej tal-Istati Uniti<br>**Evaluated Generation (Eng MT*):**<br>The Vice President is the leader of the United States. | 0.75 | 0.99 | 0.42 | 0.67 | 3rd |
| **Graph:**<br>Turkey \| demonym \| Turk<br>**Evaluated Generation (Mlt):**<br>Id-demonimu tal-abitanti tal-belt ta' Turkmen huwa ta<br>**Evaluated Generation (Eng MT*):**<br>The demonym of the inhabitants of the city of Turkmen is | 0.08 | 0.84 | 0.08 | 0.71 | 5th |

Tab. C.1: Referenceless metric evaluation on Maltese examples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (best), 3rd (medium) and 5th (worst) quintile (Q) based on the accuracy of MonoLR (mLR) compared to the human annotation (H). *The MT model may have altered differences in the original Maltese generation.

| Sample | Precision H | Precision mLR | Recall H | Recall mLR | Q |
|---|---|---|---|---|---|
| **Graph:**<br>(66063) 1998 RO1 \| meanTemperature \| 265.0 (kelvins)<br>(66063) 1998 RO1 \| apoapsis \| 254989570.60815 (kilometres)<br>(66063) 1998 RO1 \| epoch \| 2013-11-04<br>(66063) 1998 RO1 \| orbitalPeriod \| 360.29 (days)<br>**Evaluated Generation (Rus):**<br>Небесное тело, известное как (66063) 1998 RO1, имеет среднюю температуру 265 Кельвинов и орбитальный период 360,29 дней. Его апоцентр - 254989570,60815 километров, а его эпоха - 13 января 2016 года.<br>**Evaluated Generation (Eng MT*):**<br>The celestial body, known as (66063) 1998 RO1, has an average temperature of 265 Kelvin and an orbital period of 360.29 days. Its apocenter is 254989570.60815 kilometers, and its epoch is January 13, 2016. | 0.88 | 0.93 | 0.88 | 0.79 | 1st |
| **Graph:**<br>School of Business and Social Sciences at the Aarhus University \| affiliation \| European University Association<br>European University Association \| headquarter \| Brussels<br>School of Business and Social Sciences at the Aarhus University \| established \| 1928<br>**Evaluated Generation (Rus):**<br>Школа бизнеса и социальных наук Орхусского университета была создана в 1928 году и входит в Ассоциацию университетов Европы, штаб-квартира которой находится в Брюссельском столичном регионе.<br>**Evaluated Generation (Eng MT*):**<br>The School of Business and Social Sciences at Aarhus University was founded in 1928 and is a member of the Association of European Universities, headquartered in the Brussels-Capital Region. | 0.88 | 0.56 | 1.00 | 0.90 | 3rd |
| **Graph:**<br>11 Diagonal Street \| location \| South Africa<br>**Evaluated Generation (Rus):**<br>Диагонал-стрит 11 находится в Южной Африке.<br>**Evaluated Generation (Eng MT*):**<br>Diagonal Street 11 is located in South Africa. | 0.12 | 1.00 | 0.12 | 1.00 | 5th |

Tab. C.2: Referenceless metric evaluation on Russian examples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (best), 3rd (medium) and 5th (worst) quintile (Q) based on the accuracy of MonoLR (mLR) compared to the human annotation (H). *The MT model may have altered differences in the original Russian generation.

# Bibliography


Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kancharla Aditya Hari, Bhavyajeet Singh, Anubhav Sharma, and Vasudeva Varma. 2023. WebNLG challenge 2023: Domain adaptive machine translation for low-resource multilingual RDF-to-text generation (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 93–94, Prague, Czech Republic. Association for Computational Linguistics.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Oshin Agarwal, 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, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. HTLM: Hyper-text pre-training and prompting of language models. In *International Conference on Learning Representations*.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Alyssa Allen, Ashley Lewis, Yi-Chien Lin, Tomiris Kaumenova, and Michael White. 2024. OSU CompLing at the GEM'24 data-to-text task. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 100–111, Tokyo, Japan. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Felipe Almeida Costa, Thiago Castro Ferreira, Adriana Pagano, and Wagner Meira. 2020. Building the first English-Brazilian Portuguese corpus for automatic post-editing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6063–6069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

M. Bacchiani and B. Roark. 2003. Unsupervised language model adaptation. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Xuefeng Bai, Linfeng Song, and Yue Zhang. 2020. Online back-parsing for AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1206–1219, Online. Association for Computational Linguistics.

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

Satanjeev Banerjee 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, Ann Arbor, Michigan. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 16–23.

Sean Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, Lynn Andrea Stein, et al. 2004. Owl web ontology language reference. *W3C recommendation*, 10(2):1–53.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

Anja Belz. 2005. Statistical generation: Three methods compared and evaluated. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

David Bergés, Roser Cantenys, Roger Creus, Oriol Domingo, and José A. R. Fonollosa. 2020. The UPC RDF-to-text system at WebNLG challenge 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 167–170, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12564–12573.

Pavel Blinov. 2020. Semantic triples verbalization with generative pre-training model. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 154–158, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Rudolf Blum. 1991. *Kallimachos: the Alexandrian Library and the origins of bibliography*. Univ of Wisconsin Press.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Thiago Castro Ferreira, Iacer Calixto, Sander Wubben, and Emiel Krahmer. 2017. Linguistic realisation as machine translation: Comparing different MT models for AMR-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 1–10, Santiago de Compostela, Spain. Association for Computational Linguistics.

Thiago Castro Ferreira, 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, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Thiago Castro Ferreira, 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, Tilburg University, The Netherlands. Association for Computational Linguistics.

Thiago Castro Ferreira, 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, Hong Kong, China. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver. Association for Computational Linguistics.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135.

Wenhu Chen, 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, Online. Association for Computational Linguistics.

Ziming Cheng, Zuchao Li, and Hai Zhao. 2022. BiBL: AMR parsing and generation with bidirectional Bayesian learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5461–5475, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cohere. 2024. Command r: Large language model for retrieval-augmented generation and tool use. https://docs.cohere.com/v2/docs/command-r. Accessed: April 29, 2025.

Simone Conia, Edoardo Barba, Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, Riccardo Orlando, Luigi Procopio, and Roberto Navigli. 2024. MOSAICo: a multilingual open-text semantically annotated interlinked corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7990–8004, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Richard E. Cullingford. 1979. Pattern-matching and inference in story understanding. *Discourse Processes*, 2(4):319–334.

Marc Damonte and Shay Cohen. 2020. Abstract meaning representation 2.0 - four translations.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Oriol Domingo, David Bergés, Roser Cantenys, Roger Creus, and José A. R. Fonollosa. 2020. Enhancing sequence-to-sequence modelling for RDF triples to natural text. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 40–47, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramón Astudillo. 2022. Inducing and using alignments for transition-based AMR parsing.

In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1086–1098, Seattle, United States. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 115–119, Jeju Island, Korea. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Mohamed Elfeki, Rui Liu, and Chad Voegele. 2025. Return of the encoder: Maximizing parameter efficiency for slms. *Preprint*, arXiv:2501.16273.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing wikidata to the linked data web. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*, pages 50–65. Springer.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Angela Fan and Claire Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016. Generation from Abstract Meaning Representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California. Association for Computational Linguistics.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zihao Fu, Bei Shi, Wai Lam, Lidong Bing, and Zhiyuan Liu. 2020. Partially-aligned data-to-text generation with distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9183–9193, Online. Association for Computational Linguistics.

Jianfeng Gao, Hisami Suzuki, and Wei Yuan. 2006. An empirical study on language model adaptation. *ACM Transactions on Asian Language Information Processing*, 5(3):209–227.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural language generation enhances human decision-making with uncertain information. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–268, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Ning Dai, Xipeng Qiu, Xiangyang Xue, David Wipf, and Zheng Zhang. 2020a. $\surd^2$: A plan-and-pretrain approach for knowledge graph-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic*

*Web (WebNLG+)*, pages 100–106, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020b. CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, 7:297–312.

Valerie Hajdik, Jan Buys, Michael Wayne Goodman, and Emily M. Bender. 2019. Neural text generation from rich semantic representations. 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 2259–2266, Minneapolis, Minnesota. Association for Computational Linguistics.

Kelvin Han and Claire Gardent. 2023. Generating and answering simple and complex questions from text and from knowledge graphs. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–304, Nusa Dua, Bali. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Ana Marasović, and Noah A. Smith. 2021. Promoting graph awareness in linearized graph-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 944–956, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2018. Quantized neural networks: Training neural networks with low precision weights and activations. *journal of machine learning research*, 18(187):1–30.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical association*, 84(406):414–420.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mayank Jobanputra and Vera Demberg. 2024. TeamSaarLST at the GEM'24 data-to-text task: Revisiting symbolic retrieval in the LLM-age. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 92–99, Tokyo, Japan. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Mihir Kale 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, Dublin, Ireland. Association for Computational Linguistics.

J. Kang, M. Coavoux, C. Lopez, and D. Schwab. 2024. The little prince amr corpus (expanded with korean and croatian).

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Zdeněk Kasner and Ondřej Dušek. 2020. Train hard, finetune easy: Multilingual denoising for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 171–176, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Robert T. Kasper. 1989. A flexible interface for linking applications to Penman's sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Maxim Kazakov, Julia Preobrazhenskaya, Ivan Bulychev, and Aleksandr Shain. 2023. WebNLG-interno: Utilizing FRED-t5 to address the RDF-to-text problem (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 67–72, Prague, Czech Republic. Association for Computational Linguistics.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. JointGT: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538, Online. Association for Computational Linguistics.

Natthawut Kertkeidkachorn and Hiroya Takamura. 2020. Text-to-text pre-training model with plan selection for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 159–166, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2017. Abstract meaning representation (amr) annotation release 2.0.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Tim O'Gorman, Martha Palmer, Nathan Schneider, and Madalina Bardocz. 2020. Abstract meaning representation (amr) annotation release 3.0.

Kevin Knight, Laura Baranescu, Claire Bonial, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, and Nathan Schneider. 2014. Abstract meaning representation (amr) annotation release 1.0.

Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. *AAAI/IAAI*, 2000:703–710.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. 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 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1503–1514, Seattle, Washington, USA. Association for Computational Linguistics.

Nalin Kumar, Saad Obaid Ul Islam, and Ondrej Dusek. 2023. Better translation + split and generate for multilingual RDF-to-text (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 73–79, Prague, Czech Republic. Association for Computational Linguistics.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *Pacific Rim International Conference on Artificial Intelligence*, pages 453–458. Springer.

Irene Langkilde and Kevin Knight. 1998a. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Irene Langkilde and Kevin Knight. 1998b. Generation that exploits corpus-based statistical knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 704–710, Montreal, Quebec, Canada. Association for Computational Linguistics.

Guy Lapalme. 2020a. The jsrealb text realizer: Organization and use cases. *CoRR*, abs/2012.15425.

Guy Lapalme. 2020b. RDFjsRealB: a symbolic approach for generating text from RDF triples. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 144–153, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Guy Lapalme. 2024. pyrealb at the GEM'24 data-to-text task: Symbolic English text generation from RDF triples. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 54–58, Tokyo, Japan. Association for Computational Linguistics.

Ora Lassila and Ralph R. Swick. 1999. Resource description framework (rdf) model and syntax.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Teven Le Scao and Claire Gardent. 2023. Joint representations of text and knowledge graphs for retrieval and evaluation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 110–122, Nusa Dua, Bali. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xintong Li, Aleksandre Maskharashvili, Symon Jory Stevens-Guille, and Michael White. 2020. Leveraging large pretrained models for WebNLG 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 117–124, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology.*

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Michela Lorandi and Anya Belz. 2023. Data-to-text generation for severely under-resourced languages with GPT-3.5: A bit of help needed from Google Translate (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 80–86, Prague, Czech Republic. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Min Ma, Michael Nirschl, Fadi Biadsy, and Shankar Kumar. 2017. Approaches for neural-network language model adaptation. In *INTERSPEECH*, pages 259–263.

Neil Macdonald. 1954. Language translation by machine-a report of the first successful trial. *Computers and automation*, 3(2):6–10.

Manuel Mager, 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, Online. Association for Computational Linguistics.

Emma Manning and Nathan Schneider. 2021. Referenceless parsing-based evaluation of AMR-to-English generation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 114–122, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abelardo Carlos Martínez Lorenzo, Marco Maru, and Roberto Navigli. 2022. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1727–1741, Dublin, Ireland. Association for Computational Linguistics.

Abelardo Carlos Martinez Lorenzo and Roberto Navigli. 2024. Efficient AMR parsing with CLAP: Compact linearization with an adaptable parser. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5578–5584, Torino, Italia. ELRA and ICCL.

Rebecca Mason and Eugene Charniak. 2014. Domain-specific image captioning. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 11–20, Ann Arbor, Michigan. Association for Computational Linguistics.

Jonathan May and Jay Priyadarshi. 2017. SemEval-2017 task 9: Abstract Meaning Representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

Ian C McIlwaine. 1997. The universal decimal classification: Some factors concerning its origins, development, and influence. *Journal of the American society for information science*, 48(4):331–339.

Kathleen R. McKeown. 1982. The text system for natural language generation: An overview. In *20th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Toronto, Ontario, Canada. Association for Computational Linguistics.

James R Meehan. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, pages 91–98.

Yan Meng and Christof Monz. 2024. Disentangling the roles of target-side transfer and regularization in multilingual machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1828–1840, St. Julian's, Malta. Association for Computational Linguistics.

Marie W Meteer. 1991. Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304.

Francois Meyer and Jan Buys. 2024. Triples-to-isiXhosa (T2X): Addressing the challenges of low-resource agglutinative data-to-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16841–16854, Torino, Italia. ELRA and ICCL.

Simon Mille, Bernd Bohnet, Leo Wanner, and Anja Belz. 2017a. Shared task proposal: Multilingual surface realization using Universal Dependency trees. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 120–123, Santiago de Compostela, Spain. Association for Computational Linguistics.

Simon Mille, Roberto Carlini, Alicia Burga, and Leo Wanner. 2017b. FORGe at SemEval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 920–923, Vancouver, Canada. Association for Computational Linguistics.

Simon Mille, Mohammed Sabry, and Anya Belz. 2024. DCU-NLG-small at the GEM'24 data-to-text task: Rule-based generation and post-processing with t5-base. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 84–91, Tokyo, Japan. Association for Computational Linguistics.

Simon Mille, Elaine Uí Dhonnchadha, Stamatia Dasiopoulou, Lauren Cassidy, Brian Davis, and Anya Belz. 2023. DCU/TCD-FORGe at WebNLG'23: Irish rules! (WegNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Paul Molins and Guy Lapalme. 2015. JSrealB: A bilingual text realizer for web programming. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 109–111, Brighton, UK. Association for Computational Linguistics.

Sebastien Montella, 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, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto,

Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Anna Nikiforovskaya and Claire Gardent. 2024. Evaluating RDF-to-text generation models for English and Russian on out of domain data. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 134–144, Tokyo, Japan. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jekaterina Novikova, 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, Saarbrücken, Germany. Association for Computational Linguistics.

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, and Valerio Basile. 2024. DipInfo-UniTo at the GEM'24 data-to-text task: Augmenting LLMs with the split-generate-aggregate pipeline. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 59–65, Tokyo, Japan. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros,

Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Chinonso Cynthia Osuji, Rudali Huidrom, Kolawole John Adebayo, Thiago Castro Ferreira, and Brian Davis. 2024. DCU-ADAPT-modPB at the GEM'24 data-to-text generation task: Model hybridisation for pipeline data-to-text natural language generation. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 66–75, Tokyo, Japan. Association for Computational Linguistics.

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

Ankur Parikh, 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, Online. Association for Computational Linguistics.

Nivranshu Pasricha, Mihael Arcan, and Paul Buitelaar. 2020. NUIG-DSI at the WebNLG+ challenge: Leveraging transfer learning for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 137–143, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2018. Bootstrapping generators from noisy data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1516–1527, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating English from Abstract Meaning Representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25, Edinburgh, UK. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, 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(140):1–67.

Virginia Ramón-Ferrer, Carlos Badenes-Olmedo, and Oscar Corcho. 2025. Spanish triple-to-text benchmark on low-resource large language models.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 506–516, Red Hook, NY, USA. Curran Associates Inc.

Michael Regan, Shira Wein, George Baker, and Emilio Monti. 2024. MASSIVE multilingual Abstract Meaning Representation: A dataset and baselines for hallucination detection. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 1–17, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Robert Dale. 2000. Building natural language generation systems.

Ehud Reiter, Chris Mellish, and John Levine. 1995. Automatic generation of technical documentation. *Applied Artificial Intelligence an International Journal*, 9(3):259–287.

Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. Enhancing AMR-to-text generation with dual graph representations. 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 3183–3194, Hong Kong, China. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Jonas Pfeiffer, Yue Zhang, and Iryna Gurevych. 2021a. Smelting gold and silver for improved multilingual AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 742–750, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021b. 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, Online. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021c. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacques Robin and Kathleen McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, 85(1-2):135–179.

Sebastian Ruder. 2022. The state of multilingual ai. *Cited on*, page 23.

Roger C Schank, Neil M Goldman, Charles J Rieger III, and Christopher Riesbeck. 1973. Margie: Memory analysis response generation, and inference on english. In *IJCAI*, volume 3, pages 255–61.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Martin Schmitt, Leonardo F. R. Ribeiro, Philipp Dufter, Iryna Gurevych, and Hinrich Schütze. 2021. Modeling graph structure via relative position for text generation from knowledge graphs. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 10–21, Mexico City, Mexico. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, 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, Online. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017b. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est.

Anastasia Shimorina, 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, Florence, Italy. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Amit Singhal. 2012. Introducing the knowledge graph: things, not strings.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2022. Exploring a POS-based two-stage approach for improving low-resource AMR-to-text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 531–538, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago A. S. Pardo. 2020. NILC at WebNLG+: Pre-trained sequence-to-sequence models on RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 131–136, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Linfeng Song, Yue Zhang, Xiaochang Peng, Zhiguo Wang, and Daniel Gildea. 2016. AMR-to-text generation as a traveling salesman problem. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2084–2089, Austin, Texas. Association for Computational Linguistics.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2023. Phylogeny-inspired soft prompts for data-to-text generation in low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–198, Nusa Dua, Bali. Association for Computational Linguistics.

William Soto Martinez, Yannick Parmentier, and Claire Gardent. 2024. Generating from AMRs into high and low-resource languages using phylogenetic knowledge and hierarchical QLoRA training (HQL). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 70–81, Tokyo, Japan. Association for Computational Linguistics.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Trung Tran and Dang Tuan Nguyen. 2020. WebNLG 2020 challenge: Semantic template mining for generating references from RDF. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 177–185, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. GTR-LSTM: A triple encoder for sentence generation from RDF data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1627–1637, Melbourne, Australia. Association for Computational Linguistics.

Francis M. Tyers. 2009. Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Chris van der Lee, 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, Dublin, Ireland. Association for Computational Linguistics.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based AMR parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qingyun Wang, Semih Yavuz, Xi Victoria Lin, Heng Ji, and Nazneen Rajani. 2021. Stage-wise fine-tuning for graph-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 16–22, Online. Association for Computational Linguistics.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282, Dublin, Ireland. Association for Computational Linguistics.

Wikimedia Foundation. 2025. List of wikipedias. [Online; accessed 3-June-2025].

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. biom bull 1 (6): 80–83.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Sam Wiseman, 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, Copenhagen, Denmark. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016a. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016b. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2021. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 896–907, Online. Association for Computational Linguistics.

Linting Xue, 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, Online. Association for Computational Linguistics.

Zixiaofan Yang, Arash Einolghozati, Hakan Inan, Keith Diedrick, Angela Fan, Pinar Donmez, and Sonal Gupta. 2020. Improving text-to-text pre-trained models for the graph-to-text task. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 107–116, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, and Orhan Firat. 2022. Examining scaling and transfer of language model architectures for machine translation. In *International Conference on Machine Learning*, pages 26176–26192. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th*

*Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Kun Zhang, Oana Balalau, and Ioana Manolescu. 2023. FactSpotter: Evaluating the factual faithfulness of graph-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10025–10042, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15590–15606, Toronto, Canada. Association for Computational Linguistics.

Giulio Zhou and Gerasimos Lampouras. 2020. WebNLG challenge 2020: Language agnostic delexicalisation for multilingual RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 186–191, Dublin, Ireland (Virtual). Association for Computational Linguistics.