MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations

Dagmar Gromann[1], **Hugo Gonçalo Oliveira**[2], **Lucia Pitarch**[3], **Elena-Simona Apostol**[4], **Jordi Bernad**[3], **Eliot Bytyçi**[5], **Chiara Cantone**[6], **Sara Carvalho**[7], **Francesca Frontini**[8], **Radovan Garabik**[9], **Jorge Gracia**[3], **Letizia Granata**[10], **Fahad Khan**[8], **Timotej Knez**[11], **Penny Labropoulou**[12], **Chaya Liebeskind**[13], **Maria Pia di Buono**[10], **Ana Ostroški Anić**[14], **Sigita Rackevičienė**[15], **Ricardo Rodrigues**[16], **Gilles Sérasset**[17], **Linas Selmistraitis**[15], **Mahammadou Sidibé**[17], **Purificação Silvano**[18], **Blerina Spahiu**[19], **Enriketa Sogutlu**[20], **Ranka Stanković**[21], **Ciprian-Octavian Truică**[4], **Giedrė Valūnaitė Oleškevičienė**[15], **Slavko Zitnik**[11], **Katerina Zdravkova**[22]

[1]University of Vienna, Austria, dagmar.gromann@gmail.com
[2]University of Coimbra, DEI, CISUC, Portugal, hroliv@dei.uc.pt
[3]University of Zaragoza, Spain {lpitarch, jbernad, jogracia}@unizar.es
[4]National University of Science and Technology Politehnica Bucharest, Romania
{elena.apostol, ciprian.truica}@upb.ro
[5]University of Prishtina, Kosovo, eliot.bytyci@uni-pr.edu
[6]University of Siena, Italy, chiara.cantone@student.unisi.it
[7]University of Aveiro-CLLC/NOVA CLUNL, Portugal, sara.carvalho@ua.pt
[8]CNR-ILC, Italy {francesca.frontini, fahad.khan}@ilc.cnr.it
[9]Slovak Academy of Sciences, Slovakia, radovan.garabik@kassiopeia.juls.savba.sk
[10]University of Naples “L’Orientale”, Italy, {l.granata4@studenti., mpdibuono@}unior.it
[11]University of Ljubljana, Slovenia, {timotej.knez, slavko.zitni}@fri.uni-lj.si
[12]Athena R.C. – ILSP, Greece, penny@athenarc.gr
[13]Jerusalem College of Technology, Israel, liebchaya@gmail.com
[14]Institute for the Croatian Language, Croatia, aostrosk@ihjj.hr
[15]Mykolas Romeris University, Lithuania, {sigita.rackeviciene, selmistraitis, gvalunaite}@mruni.eu
[16]Polytechnic Institute of Coimbra / CISUC, Portugal, rmanuel@dei.uc.pt
[17]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France, gilles.serasset@imag.fr
[18]University of Porto / CLUP, Portugal, msilvano.letras.up.pt
[19]University of Milan-Bicocca, Italy, blerina.spahiu@unimib.it
[20]University College Beder, Albania, esogutlu@beder.edu.al
[21]University of Belgrade, Serbia ranka.stankovic@rgf.bg.ac.rs
[22]University Ss Cyril and Methodius in Skopje, North Macedonia, katerina.zdravkova@finki.ukim.mk

Abstract

Understanding the relation between the meanings of words is an important part of comprehending natural language. Prior work has either focused on analysing lexical semantic relations in word embeddings or probing pretrained language models (PLMs), with some exceptions. Given the rarity of highly multilingual benchmarks, it is unclear to what extent PLMs capture relational knowledge and are able to transfer it across languages. To start addressing this question, we propose MultiLexBATS, a multilingual parallel dataset of lexical semantic relations adapted from BATS in 15 languages including low-resource languages, such as Bambara, Lithuanian, and Albanian. As experiment on cross-lingual transfer of relational knowledge, we test the PLMs’ ability to (1) capture analogies across languages, and (2) predict translation targets. We find considerable differences across relation types and languages with a clear preference for hypernymy and antonymy as well as romance languages.

Keywords: Lexical Semantic Relations, Multilingual Benchmark, BATS

1. Introduction

A popular benchmark for exploring the capability of word embeddings and PLMs to capture lexical semantic relational knowledge are analogy datasets, such as the Bigger Analogy Test Set (BATS) (Gladkova et al., 2016) or Google Analogy Test Set (GATS) (Mikolov et al., 2013). For instance, Rezaee and Camacho-Collados (2022) probe the ability of PLMs to distinguish types, relations, and the directionality of a given relation in English by utilising BATS. Other multilingual approaches (e.g. Ulčar et al., 2020; Peng et al., 2022) focus on a multilingual comparison of static rather than contextual embeddings or PLMs. A reliable comparison across languages is rare for embeddings as well as PLMs due to a lack of reliable multilingual benchmarks. In order to provide a richer benchmark, we propose Multilingual Lexicographic BATS (MultiLexBATS), a highly multilingual dataset of lexical relations adapted from BATS in 15 languages spanning language families from Romance and Balto-Slavic to Mande[1].

We argue that to allow for reliable comparisons across languages and language families, a carefully curated and parallel, aligned dataset of lexical semantic relations is required. This can best be achieved by means of manual translations by first-language speakers in order to control the quality of the resulting dataset. Since the relation targets in analogies might comprise more than one correct answer, we ruled out datasets with this limitation, e.g. GATS (Mikolov et al., 2013) or Ulčar et al. (2020). Instead, we decided to opt for adapting the lexical semantic relation pairs of BATS in 15 languages.

Mickus et al. (2023) derive a Multilingual Analogy Test Set (MATS) by adapting BATS in Dutch, French, German, Italian, Mandarin, and Spanish, however, without multi-word expressions or alignment across languages. BATS has also been adapted to Icelandic (Friðriksdóttir et al., 2022) and Japanese (Karpinska et al., 2018). These datasets, just like Ulčar et al. (2020), dropped multi-word expressions and are not or only partially directly aligned across languages, which naturally restricts the kind of potential cross-lingual tasks and experiments. As a novel cross-lingual analogy-based probing task, we introduce the prediction of translation targets and compose cross-lingual prompts of the type *apple is to fruit (en) as manzana es como ... (es)*, where the PLM should provide *fruta* as an answer.

Aside from the manually curated, fully aligned dataset of lexical semantic relations in 15 languages, including languages of different morphological richness, alphabets, and writing systems (right-to-left in Hebrew), this paper makes the following additional contributions:

- analogy templates in all languages, e.g. *<a> është për <b> ashtu si <c> për <d>* (Albanian);
- translation guidelines of how to handle missing or impossible translations, e.g. the *50/50 burger* of half ground bacon and half ground beef patty;
- a highly multilingual lexical-semantic relation dataset with multi-word expressions;
- a cross-lingual translation experiment based on language-aligned analogies.

The major contribution is the aligned multilingual dataset itself and the experiments serve the purpose to inspect how challenging analogy-based tasks in the different languages of the proposed dataset are. Multi-word expressions are particularly interesting, since they propose novel challenges to the classical analogy prediction task with only single words. Furthermore, we decided not to delete cases where we observed issues in the English dataset, but instead mark and keep them for further analysis across languages. Thereby, it is possible to detect if an English word has no translation equivalence in other languages.

2. Related Work

Since this paper proposes a dataset and corresponding experiments with neural language models, the presentation of related work is structured accordingly.

Datasets As one of the first analogy datasets, the Google Analogy Test Set (Mikolov et al., 2013) has been widely used and cited. However, as set out in Gladkova et al. (2016), it suffers from some weaknesses that BATS seeks to overcome. For instance, BATS (Gladkova et al., 2016) allows for multiple valid answers to the analogy task, whereas GATS only permits one. The two datasets that are probably closest to the proposed MultiLexBATS in terms of language coverage are Multilingual Analogy Test Set (MATS) (Mickus et al., 2023) and the one proposed by Ulčar et al. (2020). Mickus et al. (2023) propose MATS by adapting BATS in Dutch, French, German, Italian, Mandarin, and Spanish, removing several multi-word expressions, analogically incorrect pairs and unidiomatic translations. While comparable in content, the individual languages are ordered alphabetically and not aligned to each other or the English original dataset. Ulčar et al. (2020) create an initial analogy dataset in Slovene that is then translated to Croatian, Estonian, Finnish, Swedish, Latvian, Lithuanian, and English by means of querying Wikipedia, Wikidata, and BabelNet with a final human quality control. The dataset consists of five semantic and ten syntactic/morphological categories, i.e., no lexical semantic relations are provided. The alignment across languages is limited to specific pairs of languages and only one valid answer per analogy is expected. Similar to MATS, the authors exclude multi-word expressions.

The CogALex shared task (Xiang et al., 2020) provided a lexical semantic relation dataset in English, German, Chinese, and Italian. Given a word pair, submitting systems were challenged to detect whether they relate per synonymy, antonymy, hypernymy or not at all (random).

[1] The dataset, detected issues in BATS, all test prompt templates, randomly selected pairs for experiments, and the code for our experiment are available at `https://github.com/nexuslinguarum/MultiLexBATS`

TALES (Gonçalo Oliveira et al., 2020) is a Portuguese dataset with the same format and goal as BATS. It was created automatically from ten Portuguese lexical resources and is focused on lexical semantic relations. Besides hypernymy, synonymy, antonymy and part-of, it covers purpose-of relations.

Experiments Analogical reasoning has been widely used to probe relational knowledge in language models. In English, analogical reasoning has been tested on both static (Mikolov et al., 2013) and contextual embeddings (Petroni et al., 2019; Bouraoui et al., 2020; Ushio et al., 2021). In English, analogies have been powerful tools to uncover abstract relations (Petroni et al., 2019; Ushio et al., 2021) and fine-graded features within the relations, such as the type or directionality (Rezaee and Camacho-Collados, 2022). Existing literature on analogical reasoning in multilingual settings rarely focuses on probing the cross-lingual abilities of PLMs. For contextual embeddings, Artetxe et al. (2016) evaluate the preservation of monolingual linguistic features while performing cross-lingual transfer tasks, Brychcín et al. (2019) use analogies to compare multi- and monolingual settings in language transfer, and Ulčar et al. (2020) compare the quality of static embeddings in different languages. In PLMs, analogies have been used to boost the global consistency in language transfer (Garneau et al., 2021) and to evaluate the performance of different training strategies (with and without global co-occurrence) (Ai and Fang, 2023). One of the few approaches to probe PLMs through analogies is presented by Mickus et al. (2023), building on the approach proposed by Petroni et al. (2019).

3. Dataset Curation Guidelines

To ensure full equivalence of translations, we assigned one ID to each set of source word and target words across relation files. In the original BATS dataset, source words are tab-separated by a list of backslash-separated target words. While this format works well monolingually, it creates considerable issues when seeking to represent (missing) equivalence across languages. Thus, our initial translation dataset is line-separated and aligned to allow for adding cross-lingual comparison as represented in Table 1.

In the end, we compose files with all languages, where each language is represented in a column aligned with English and all other languages. Thereby, we are able to explicitly indicate (missing) equivalences of target words as well as duplication of equivalent words in the target language. Furthermore, this way of structuring the data enables cross-lingual and multilingual comparisons and experiments on the dataset as exemplified in Section 6. For the translation process, we introduced the following labels:

- DUPLICATE_target_word: if a target word has the exact same translation as a previous target word associated with the same relation to the same source word, use this label and replace ”target_word” in the label with the language-specific duplicated occurrence;
- NO_TRANSLATION: there is no translation for the target word in the specific context of this relation to a specific source word, e.g. there is a general translation for *gun* but not in the context of a car;
- COMMENT: any additional observation regarding the English dataset or the translation task.

For each DUPLICATE or NO_TRANSLATION, we try to provide an alternative target word in the target language to keep the count of target words equivalent to the original dataset. This is not always possible since for some source words, e.g. for *allosaurus*, there is only a limited number of hypernyms.

4. Handling Detected Issues

This task of translating the original BATS lexical semantic relations part uncovered a substantial number of issues in the English dataset. These issues relate to (1) source word being identical to a target word in the same set, e.g. *bird* (source word) cannot be meronymically related to *bird* (target word), (2) duplicated target words, e.g. a *pony* is only one time a *mammal*, (3) errors in target words, e.g. *physical physical entity*, (4) wrong source-target pairs, e.g. *lemon* (source word) is a *citrus fruit* (target word) but not a type of *garden truck* (target) and *fox* is not a hyponym of *domestic animal*, (5) incomplete multi-word sequences, e.g. *picture* relates to *book* as *picture book*, (6) polysemous source-target sets, e.g. *notebook* is associated with two sets of semantically grouped target words: (a) words related to the sense of *book*, (b) words related to the sense of *computer*, and (7) mixed part-of-speech (POS) tags in a set, e.g. *coyote* (noun) is related to *placental* (adjective).

In the case of (1), we decided to remove the target word that is identical to the source word. For (2), we marked them as DUPLICATE and removed duplicate relations centrally during the experiments. In the case of (3), most cases are erroneous duplicates of already existing target words, that is, there is *physical physical entity*

ID	Relation	Source Words	Target Words	DE	Comments
L06_7	meronyms - part	car		Auto	-
			engine	Motor	-
			horn	Hupe	-
			hooter	DUPLICATE_Hupe	-
			trunk	Kofferraum	-
			gun	NO_TRANSLATION	unclear link to car
			armrest	Armlehne	-
			...	..	-

Table 1: MultiLexBATS example of the dataset organization with labels and comments

as well as *physical entity*, which is why we removed these cases. For (4), we marked them as NO_TRANSLATION in the target language and tried to provide an alternative target word. The case of (5) is particularly tricky, for which we decided to add the missing compound word based on our joint best guess and translate the words correspondingly. This is due to the fact that BATS has mostly been generated automatically, where compounds have been split and added as two target words or at times even source and target word, e.g. *bed* (source) and *twin* (target) which is a *twin bed*. For now, in the case of (6), polysemous sets, we simply translated them as two separate sets and kept them in the dataset. The switching between POS tags (7) within source-target sets and overall the identification of the correct POS without any other context than the other words in the set and the type of relation was particularly challenging and we translated according to our joint best guess. For better traceability, we publish a list of identified issues alongside the dataset.

In any case, multi-words have to be considered since many translation equivalents are naturally multi-word even if the original English is a single word. Thus, we explicitly also included the multi-word expressions in the original dataset in English, contrary to previous works. This poses some challenges to existing probing and training approaches, however, we believe that it is more useful than unnaturally limiting the dataset to single words only.

5. Dataset Description

The final dataset consists of 15 natural languages that are aligned based on the English version. The relations are equivalent to BATS animal and miscellaneous hypernyms, miscellaneous hyponyms, meronyms of substances, members, and parts, exact synonyms and of intensity, and exact, gradable and binary antonyms. In this section, we present the classification of these languages by language family as well as detailed statistics on the dataset for each language.

5.1. Language Families in MultiLexBATS

In this joint translation endeavour, the final MultiLexBATS dataset represents 15 natural languages from three major language families and eight subfamilies, such as Baltic, Slavic, Romance, and Manding as depicted in Fig. 5.1. We classified the languages based on Glottolog (Hammarström et al., 2023). The dataset covers a wide range of different linguistic features, such as degree of inflection, as well as different alphabets, such as modern Greek, and even writing systems, such as right-to-left in modern Hebrew.

5.2. MultiLexBATS Statistics

The original BATS dataset of lexicographic relations consists of 50 source words per relation and a varying number of associated target words. Not all BATS words were translated in MultiLexBATS due to various reasons, such as lexical gaps (NO_TRANSLATION) and repetitions in the original BATS dataset as well as duplicates due to identical translation equivalents (DUPLICATE). In both cases, translators tried to provide alternative target words for the specific source word. The final statistics of the resulting source and associated target words per language are represented in Table 2, including the Fleiss κ inter-annotator agreement for all languages with more than one translator. For this agreement, 20 randomly selected sets of source-target words per relation were translated by each translator and the κ values are at the upper end of moderate to almost perfect.

The number of alternative target words provided varies significantly across languages, but, even if not as much, so do other figures. We note that two languages were not completed, and have several missing translations, namely Bambara (BM) and Macedonian (MK), thus the lower number of source words.

6. Experiments

The original idea of the analogy completion task was to test relational knowledge in vector spaces

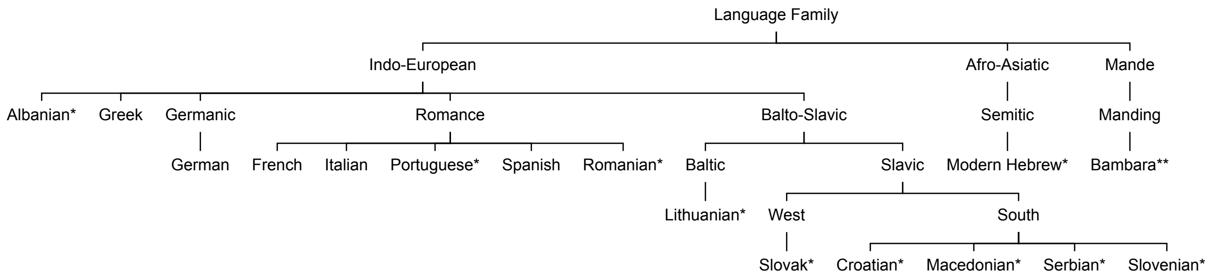

Figure 1: MultiLexBATS Languages by family according to Glottolog; languages that we consider low-resource in the cross-lingual transfer are marked with *

Lang	Transl	κ	Sources	Targets	NoTrans	Dupl	Altern
AL	3	0.51	500	4665	62	1042	0
BM	1	–	451	3355	174	1156	32
DE	1	–	500	5159	36	550	582
EL	1	–	500	4722	78	966	248
ES	3	0.55	500	5184	58	525	405
FR	1	–	500	4905	10	778	19
HE	1	–	500	4676	11	1027	130
HR	1	–	500	4382	98	1325	219
IT	2	0.50	500	4807	26	902	1354
LT	3	0.77	500	4506	44	1206	552
MK	1	–	423	2414	122	635	44
PT	4	0.62	500	4572	50	1136	451
RO	2	0.97	500	5054	63	650	0
SK	1	–	500	4395	225	1304	477
SL	2	0.48	500	4775	54	932	38

Table 2: Figures on MultiLexBATS per covered language.

and language models, i.e., test the ability to identify/predict targets of a relation. It is particularly interesting to test this type of knowledge across languages, which we propose with the experiments presented in this section. The baseline for evaluating the proposed dataset is a comparison to the original English BATS dataset as well as the multilingual MATS dataset and validation is further provided by the inter-annotator reliability in case of more than one translator. Nevertheless, lower performance in a language indicates that specific languages or language families are particularly challenging for multilingual language models, which in our experiments varies across languages. We test masked language models as well as one multilingual generative pre-trained transformer called BLOOM (Scao et al., 2022). In addition to the analogy completion task, we present a translation prediction task.

6.1. Analogy Completion Task

As presented by Mikolov et al. (2013), the goal of analogy completion is to predict a target word d that is related to a source word c, given a pair of source and target words a and b that express the same relation (e.g. hypernyms: $a =$*anaconda*, $b =$*snake*, $c =$*ant*, $d =$*insect*). Analogy-solving is thus the most intermediate task that MultiLexBATS can be used for, in this case, in all the provided languages.

Among other methods, the most popular approach to this task has been to compute the vector offset in a model of distributional similarity, e.g. $\vec{d} = \vec{b} - \vec{a} + \vec{c}$. More recent approaches rely on prompting language models for mask-filling or text generation. We tested both approaches in MultiLexBATS, using the prompt templates in Table 3. For German, Spanish, French and Italian, we utilised the same templates as proposed in MATS for comparability. The other templates were suggested by first-language speakers of the respective languages and are all translations or variations of the English prompt. In our GitHub, all tested templates for each language are reported, whereas in Table 3 we only report prompt templates that performed best and were finally used in the experiments.

6.2. Analogy Completion with Masked Language Models

It is generally interesting to test the performance of masked language models on analogy comple-

Language	Prompt
EN	“<a>” is to “<b>” as “<c>” is to “<d>”.
AL	“<a>” është për “<b>” ashtu si “<c>” për “<d>”.
BM	“<a>” ye “<b>” ye i n’a fɔ “<c>” ye “<d>”
DE	“<a>” ist so zu “<b>” wie “<c>” zu <d> ist.
EL	το “<a>” είναι προς το “<b>” ό,τι το “<c>” προς το “<d>”.
ES	“<a>” es a“<b>”como “<c>” es a “<d>”.
FR	“<a>” est à “<b>” ce que “<c>” est à “<d>”.
HE	“<a>” ל “<b>” כמו “<c>” ל “<d>”
HR	Odnos između riječi “<a>” i “<b>” jednak je odnosu između riječi “<c>” i “<d>”.
IT	“<a>” sta a “<b>” come “<c>” sta a “<d>”.
LT	“<a>” yra “<b>” taip, kaip “<c>” yra “<d>”.
MK	Односот меѓу зборовите “<a>” и “<b>” е еднаков со односот меѓу зборовите “<c>” и “<d>”.
PT	“<a>” está para “<b>” assim como “<c>” está para “<d>”-
RO	“<a>” este pentru “<b>” cum “<c>” este pentru “<d>”.
SK	Slovo “<a>” sa má k slovu “<b>” ako slovo “<c>” k slovu “<d>”.
SL	Beseda “<a>” je besedi “<b>” enako, kot je beseda “<c>” besedi “<d>”.

Table 3: Language-specific prompt templates for analogy completion in MultiLexBATS

tion, especially multilingual ones. The authors of MATS (Mickus et al., 2023) test a prompt-based approach with a multilingual BERT (Devlin et al., 2019) model (mBERT). For the languages shared by both datasets, we repeat the previously conducted experiment with MultiLexBATS, using the code provided in the MATS repository[2].

Due to the large number of languages, we reduced the number of computed analogies. We opted for: (1) using only prompts with quotes around the source and target words, which led to the best results in MATS; (2) computing only three analogies per source word. Instead of $50 \times 49 = 2,450$, this resulted in $50 \times 3 = 150$ analogies per relation. The three pairs used for the first part of the analogy were always selected randomly across the 49 pairs of the same relation. Diversely, while, in MATS, the authors use only the mBERT model and zero-shot, we also tested with the XLM-R (Conneau et al., 2020) model and run the experiments both in zero- and few-shot scenarios in MultiLexBats. Both models were pre-trained in more than 100 languages, including all of those in our dataset, except Bambara.

In order to complete the analogies, masked language models are prompted with the templates in Table 3, where <a> and <b> are replaced by a source, target pair, <c> is replaced by another source, and <d> is replaced by a mask token, to be predicted by the model. The analogy is considered successfully completed by the model if it predicts one of the target words associated with the source word <c>. If the target is a multi-word expression or multi-token word, a set of prompts is considered, one with each possible number of tokens in target words associated with source <c>. As long as the prediction is contained in the set of target words, the analogy is considered to be completed correctly. For the few-shot scenario, a sequence of five complete prompts (five-shot) are concatenated. Pairs of words in the prompts are randomly selected from all analogies of the same subcategory not including the source word <c>. In order to test this in a few-shot scenario, the prompts were concatenated to a sequence of five complete prompts (five-shot), randomly selected from all analogies of the same subcategory not including the source word <c>.

Average accuracies obtained in this experiment are reported in Table 4 and directly compared to results on the previously proposed multilingual dataset MATS. Performance is variable across languages, which is in line with previous results and the fact that analogy completion is a challenging task. Differences between models are not substantial, but the gains of a five-shot approach are clear.

For the configuration that more frequently achieved the best performance, five-shot in mBERT, Table 5 discriminates the accuracy by relation type. It becomes clear that most languages perform relatively well in the first two relations, hypernyms-animals (HA) and hypernyms-miscellaneous (HM), while performing poorly in meronymy, synonymy, and antonymy.

	mBERT		**XLM-R**		**mBERT**
	(MultiLexBATS)		(MultiLexBATS)		(MATS)
Lang	0-shot	5-shot	0-shot	5-shot	0-shot
DE	0.196	0.215	0.202	**0.217**	0.186
EN	0.214	0.233	0.183	**0.243**	0.233
ES	0.186	**0.236**	0.176	0.188	0.170
FR	0.223	0.219	0.221	**0.254**	0.208
IT	0.203	**0.260**	0.119	0.166	0.178

Table 4: Average accuracy in analogy completion with masked language modelling. Results for quoted templates for MultiLexBATS and MATS.

6.3. Analogy Completion with a Generative Model

To test the performance of generative pre-trained transformers on the proposed task of predicting relations in a highly multilingual analogy completion task, we utilise the BLOOM (Scao et al., 2022) model via the HuggingFace Interface API[3]. The reason for opting for this model is that it represents a freely available model based on a collaborative, research initiative, which we believe to be an important point and an excellent reason for favouring

[2] https://github.com/ATILF-UMR7118/MATS

[3] https://huggingface.co/inference-api

Lang	HA	HM	HOM	MS	MM	MP	SI	SE	AG	AB
DE	0.86	0.79	0.06	0.12	0.07	0.07	0.06	0.06	0.06	0.02
EM	0.85	0.89	0.09	0.20	0.05	0.09	0.02	0.08	0.07	0.09
ES	0.76	0.91	0.04	0.04	0.01	0.01	0.00	0.01	0.04	0.06
FR	0.88	0.90	0.15	0.17	0.03	0.16	0.04	0.11	0.04	0.07
IT	0.71	0.52	0.06	0.15	0.02	0.05	0.01	0.03	0.03	0.09

Table 5: Accuracy in analogy completion with mBERT in a 5-shot learning scenario; relation types are hypernyms-animals (HA), hypernyms-miscellaneous (HM), hyponyms-miscellaneous(HOM), meronyms-substance (MS), meronyms-member(MM), meronyms-part (MP), synonyms-intensity (SI), synonyms-exact (SE), antonyms-gradable (AG), antonyms-binary (AB).

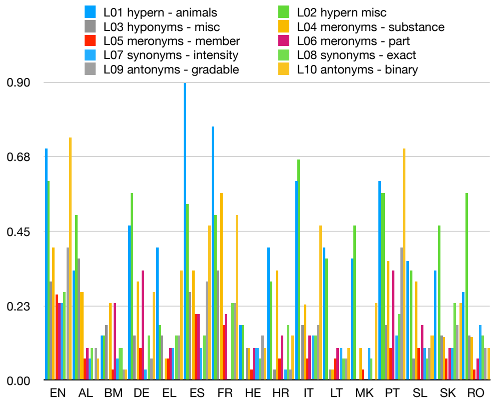

Figure 2: Results from running 30 random analogy pairs on BLOOM

BLOOM over paid alternatives[4]. The objective of this task is to evaluate cross-lingual differences in the capability of automatically acquiring lexical semantic relations from generative language models to see which languages are most challenging, the results of which are represented in Figure 2 and the detailed numbers are reported on our GitHub. The experiment relies on randomly composing 30 unique analogies that exist across all languages without any other markers, such as DUPLICATES. We provide the filled prompt templates as well as the relation target set for each language and relation on our GitHub for easier reproducibility with the 30 random prompts. The validity of the experiment and dataset is provided by the comparison to the original English BATS dataset and the multilingual MATS results as well as the inter-annotator reliability.

From this experiment and from Figure 2, it becomes clear that some relations are more challenging than others, e.g. animal hypernyms always outperform other relations, especially the worst performing member meronyms and intensity synonyms across all languages. The best overall performance was achieved by Portuguese (0.36) followed by French (0.35), Spanish (0.34), and Italian (0.28). Slavic languages performed around 0.20, led by Slovak and Macedonian. German achieved 0.24 on average. Also for this experiment, Bambara (0.12), Greek (0.16), and Hebrew (0.11) were the most challenging languages and also in Lithuanian (0.13) the model struggled with analogy completion. This is in line with our assumption that low-resource and highly inflected languages as well as different alphabets and writing systems are harder to handle for the models than high-resource languages.

For comparison purposes, we ran the approaches based on masked language modelling in the same set of 30 analogies of each relation. Average accuracies are reported in Table 6. For Macedonian, two relations (HO, AG) were excluded due to the fact that they could not be completed in time.

	mBERT		XLM-R		BLOOM
Lang	0shot	5shot	0shot	5shot	0shot
AL	0.08	0.15	0.10	**0.22**	0.20
BM	0.08	**0.30**	0.07	0.26	0.12
DE	0.03	0.10	0.03	0.09	**0.24**
EL	0.00	0.00	0.10	**0.18**	0.16
EN	0.17	0.36	0.12	0.25	**0.41**
ES	0.16	0.29	0.13	0.28	**0.34**
FR	0.20	0.29	0.19	**0.41**	0.35
HE	0.08	0.21	0.14	**0.24**	0.11
HR	0.20	0.45	0.15	**0.41**	0.16
IT	0.18	**0.39**	0.11	0.26	0.28
LT	0.05	**0.24**	0.08	**0.24**	0.13
*MK	0.18	**0.39**	0.13	0.20	0.20
PT	0.16	0.33	0.12	0.27	**0.36**
RO	0.10	**0.26**	0.11	0.31	0.17
SK	0.11	0.21	0.13	**0.28**	0.20
SL	0.11	**0.28**	0.14	0.26	0.17

Table 6: Average accuracy in analogy completion with masked language models and with BLOOM.

For each language, best performances are shared by BLOOM, mBERT and XLM-R. However, the masked language models only achieve their top performance in this task with few-shot learning, which was not performed with BLOOM and is left for future work. It is interesting to note that

[4] Several test runs on some freely available, alternative models, such as mGPT or GPT4All, returned worse results

the overall performance is higher for languages included in the pre-training dataset of BLOOM, which, in order of size, are English, French, Spanish, and Portuguese.

We also note the performance of Greek and Hebrew in mBERT with zero-shot and in BLOOM, which is comparatively low. This could potentially be attributed to the different alphabet used by these languages and the smaller size of mBERT's vocabulary in general (110k tokens vs 250k of XML-R and BLOOM). For BLOOM, the explanation might be attributed to difficulty of grasping the language given the rather short input sequence of an analogy.

6.4. Analogy-Based Translation

In order to gain deeper insights into the ability of language models to perform cross-lingual transfer based on relational knowledge, we propose a new task: analogy-based translation completion. The main difference towards the previous experiments is that the prompt consists of an analogy template in two languages, i.e., with the first part in one language and the second in another. An example for English to Spanish would be: *apple is to fruit as manzana es como ...*, where the model should provide *fruta*.

We highlight that this can be tested in MultiLexBATS due to its parallel nature, i.e., all source and target words are aligned indirectly, through English. This is different from MATS, where relation files have source words in alphabetical order and there is no explicit link between words in different languages. In fact, many existing relation datasets that do not focus on entity-relation pairs have a missing alignment across languages.

For this, we tested masked language models only. Since evaluating all possible language combinations on this task goes beyond the scope of this paper, pairs of language were selected based on the following considerations: high-resource to high-resource language (EN→DE, EN→FR, FR→EN); languages of the same family, going from high-resource to low-resource FR→RO, ES→PT, ES→IT, HR→SL, HR→SK); high-resource to low-resource language of different families, including the least resourced (EN→BM, EN→AL) and with a different alphabet (EN→EL, EN→HE). The scenario with the highest potential to support digital language equality is the transfer from high-resource to low-resource languages, where the goal is to leverage higher-resource languages in order to obtain knowledge in a lower-resource one.

Table 7 depicts the accuracies achieved with mBERT and XML-R, in zero and five-shot scenarios. They are computed for all <a>, <b>, <c>, <d> tuples in MultiLexBATS such that <a> and <b> are words in the first language and <c>, <d> are in the second, after the removal of duplicates and instances of no translation.

	mBERT		**XLM-R**	
L1→L2	0-shot	5-shot	0-shot	5-shot
EN→DE	0.064	**0.092**	0.046	0.058
EN→FR	**0.125**	0.116	0.051	0.080
FR→EN	**0.188**	0.170	0.046	0.081
FR→RO	0.073	**0.099**	0.029	0.050
ES→PT	0.117	**0.149**	0.068	0.119
ES→IT	0.070	**0.156**	0.027	0.040
SK→HR	0.074	**0.126**	0.052	0.050
HR→SL	0.161	**0.224**	0.124	0.141
HR→SK	0.081	**0.123**	0.065	0.105
EN→BM	0.007	**0.030**	0.002	0.019
EN→AL	0.021	0.047	0.019	**0.048**
EN→EL	0.001	0.001	0.019	**0.032**
EN→HE	0.007	**0.033**	0.005	0.025
FR→BM	0.005	**0.034**	0.002	0.019

Table 7: Accuracy of analogy-based translation

Given that this task is more difficult than the classical analogy completion task, it can be expected that the performance is correspondingly lower. On top of this, only a single answer was possible, i.e., the target d, aligned with b, instead of words from a set of targets. This was confirmed by the low performance on this task, with an accuracy that is rarely higher than 10%. A notable exception was the transfer from Croatian to Slovene, where the mBERT accuracy outperformed the accuracy of Slovenian in the classical analogy, also showing the proximity of both languages.

Furthermore, performance is generally better for languages of the same family. For them, the best performance was always achieved with mBERT in a five-shot scenario. The poorest performance, sometimes close to zero, was achieved when leveraging on English to obtain translations in low-resource languages.

7. Discussion

The proposed dataset and experiments clearly show that specific languages are easier to handle for the models than others, especially languages from the romance family. There are also clear performance differences across relations. The tested models clearly perform better on animal hypernyms than the most challenging meronyms and synonyms. These performance differences cannot be attributed to the language data or translation quality, since in comparison to BATS translations in languages that overlap with the languages proposed herein, our dataset leads to better results on identical tasks. In order to be able to compare directly to previous experiments with similar datasets, we opted for the simple binary evaluation with accuracy. In future work, it would be interesting to experiment with a more informative evaluation metric. Analogy and translation completion are challenging tasks in a multilingual and especially cross-lingual setting.

For the analogy-based translation, we required translations of the analogy templates to the respective languages. While this is a straightforward process for most languages we cover, in highly inflected languages, the exact translation often requires changing the case of nouns and/or adjectives, or explicit orthographic alternation in the adjacent context. For instance, the exact translation of <*man*> *is to* <*woman*> would be <*muškarac*> *je* <*ženi*> in Croatian, changing the noun <*žena*> to dative. Since automatically applying these changes to the required items in the dataset is not feasible, especially not with multi-word expressions, we opted for using a slightly adapted analogy template reformulating the "is to" to, e.g. "is related to".

In terms of lessons learned, the translation process turned out to be highly challenging given that we only had the source word, other target words, and the type of relation as a context to determine the correct translation. Furthermore, there are several scientific words in the dataset for animals and plants for which we mostly resorted to Latin names. Several words are highly culture-specific and inevitably lead to lexical gaps, such as *mourning ring*. Others possess connotations that might not exist in other cultures, such as *chuckle* or *dislike* and their synonyms, since the degree to publicly express emotions varies across cultures. Given the issues in the English dataset that we reported in Section 4, the process was further complicated. With the substantial number of issues, especially incomplete compounds, it is likely that there are some inconsistencies across languages on how these issues were treated. The task as such is difficult for humans, since there are many different choices to translate the same words, for instance, lexical variants and orthographic variants. Furthermore, differences of translator profiles, such as background, experience, and age group, might influence the choice of translation. For instance, the many slang and informal words were translated differently by distinct age groups. While all of these variations are valid translations, these differences lead to a lower kappa score. Nevertheless, we believe that this dataset represents a valuable contribution to the task of multilingual and cross-lingual relation probing and analogical reasoning as well as cross-lingual transfer experiments.

In both conducted experiments it is interesting to observe that the generative pre-trained transformer has a tendency to predict Asian characters only or with words from the input language. While the potential of a translation completion task is very high to contribute to generating language resources, the current performance leaves ample room for improvement. In fact, even with few-shot prompts, the cross-lingual transfer in masked language models is very low. It is likely that the performance might increase drastically with fine-tuning or other methods of adapting the models to this particular task, however, our intention was to probe the cross-lingual relational knowledge inherent in pre-trained language models.

8. Conclusion

The MultLexBATS dataset provides a playground for computational approaches to multilingual and cross-lingual tasks that can benefit from relational knowledge, which we tested with analogy completion. Since it is aligned across all 15 languages it equally allows for translation-oriented experiments, such as the translation completion task building on multilingual analogy template combinations. By detecting and handling apparent issues in the original English dataset, such as misspellings and duplicates, the provided dataset represents a cleaned version. However, we expected languages with similarity in morphology/language family to lead to more similar results in the zero- and few-shot experiments of cross-lingual transfer in this task. Nevertheless, the conducted experiments provide some baselines and examples of the type of tasks that can be achieved with the proposed dataset.

As part of our future work, we intend to focus on the errors to understand the intricate details of the difficulties of the tasks and differences between models. Furthermore, since this is a translation-based dataset, we are interested in analysing lexical gaps across languages. We hope that many further experiments with different types of LLMs and other approaches on lexical semantic relations will be conducted on the proposed dataset, which can use our experiments as a baseline.

9. Acknowledgements

This paper was based upon work in COST Action CA18209 NexusLinguarum, supported by COST (European Cooperation in Science and Technology). http://www.cost.eu/. Hugo Gonçalo Oliveira and Ricardo Rodrigues are partially supported by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020). C.O. Truică and E.S. Apostol are supported in part by project DECIP ("Dezvoltarea Capacitatii Institutionale a Universitatii POLITEHNICA din Bucuresti", contract no. PFE57/2022) and by the National University of Science and Technology Politehnica Bucharest through the PubArt program. This paper has been also partially supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

10. Bibliographical References

Xi Ai and Bin Fang. 2023. Multilingual pre-training with self-supervision from global co-occurrence information. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7526–7543, Toronto, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463.

Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. 2019. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*, 135:287–295.

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.

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.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, Steinþór Steingrímsson, and Einar Sigurdsson. 2022. IceBATS: An Icelandic adaptation of the bigger analogy test set. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4227–4234, Marseille, France. European Language Resources Association.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2021. Analogy training multilingual encoders. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12884–12892. AAAI Press.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Hugo Gonçalo Oliveira, Tiago Sousa, and Ana Alves. 2020. TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In *Proceedings of the ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020)*, volume 2693 of *CEUR Workshop Proceedings*, pages 41–47. CEUR-WS.org.

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.

Timothee Mickus, Eduardo Calò, Léo Jacqmin, Denis Paperno, and Mathieu Constant. 2023. „mann" is to "donna" as 「国王」 is to « reine » adapting the analogy task for multilingual and contextual embeddings. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 270–283, Toronto, Canada. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR), Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Xutan Peng, Mark Stevenson, Chenghua Lin, and Chen Li. 2022. Understanding linearity of cross-lingual word embedding mappings. *Transactions on Machine Learning Research*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Kiamehr Rezaee and Jose Camacho-Collados. 2022. Probing relational knowledge in language models via word analogies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3930–3936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? 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 3609–3624, Online. Association for Computational Linguistics.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex shared task on monolingual and multilingual identification of semantic relations. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 46–53, Online. Association for Computational Linguistics.

11. Language Resource References

Gladkova, Anna and Drozd, Aleksandr and Matsuoka, Satoshi. 2016. *Bigger Analogy Test Set*. Association for Computational Linguistics.

Harald Hammarström and Robert Forkel and Martin Haspelmath and Sebastian Bank. 2023. *Glottolog 4.8.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Mickus, Timothee and Calò, Eduardo and Jacqmin, Léo and Paperno, Denis and Constant, Mathieu. 2023. *Multilingual Analogy Test Set*. Association for Computational Linguistics.