„Mann“ is to “Donna” as「国王」is to « Reine »
Adapting the Analogy Task for Multilingual and Contextual Embeddings

Timothee Mickus[♠] **Eduardo Calò**[♡] **Léo Jacqmin**[♢]
Denis Paperno[♡] **Mathieu Constant**[♣]

[♠]University of Helsinki, timothee.mickus@helsinki.fi
[♡]Utrecht University, {e.calo,d.paperno}@uu.nl
[♢]Orange Labs, leo.jacqmin@orange.com
[♣]ATILF, CNRS/Université de Lorraine, mconstant@atilf.fr

Abstract

How does the word analogy task fit in the modern NLP landscape? Given the rarity of comparable multilingual benchmarks and the lack of a consensual evaluation protocol for contextual models, this remains an open question. In this paper, we introduce MATS: a multilingual analogy dataset, covering forty analogical relations in six languages, and evaluate human as well as static and contextual embedding performances on the task. We find that not all analogical relations are equally straightforward for humans, static models remain competitive with contextual embeddings, and optimal settings vary across languages and analogical relations. Several key challenges remain, including creating benchmarks that align with human reasoning and understanding what drives differences across methodologies.

https://github.com/ATILF-UMR7118/MATS

1 Introduction

Ever since the work of Mikolov et al. (2013b), analogy solving has been a staple of public outreach in NLP: It has been featured both in science communication[1] and in the classroom.[2] This task consists in finding a target word b_2, given a cue word b_1 it is related to, and another pair of words a_1 and b_2 that express the same relation. For example, we can ask what is the word that relates to “*king*” in the same manner that “*woman*” relates to “*man*”: This target ought to be “*queen*”.

The introduction of pre-trained contextualized embeddings (Peters et al., 2018) opened up a new research area where to expand prior knowledge about static models. This includes the analogy task. Suggestions have been put forward as to how to best adapt it: Ushio et al. (2021) propose to use a prompt-based method, whereas Vulić et al. (2020) and Lenci et al. (2022) try to derive static embeddings from BERT to fall back on the algorithm of Mikolov et al. (2013b). However, much work remains to be done to properly contrast and compare the performance of contextual and static embedding models on the analogy task. Another observation to be made is that reliable comparisons across languages are rare. On the one hand, datasets for English—such as the GATS (Google Analogy Test Set, Mikolov et al., 2013a) and BATS (Balanced Analogy Test Set, Gladkova et al., 2016) benchmarks—have been adapted or translated to a wide variety of languages. On the other hand, approaches specifically focusing on establishing multilingual comparisons are, to our knowledge, limited to Grave et al. (2018), Ulčar et al. (2020), and Peng et al. (2022)—none of which considers contextual embeddings.

How do embeddings—and in particular contextual models—perform on the analogy task beyond English? In the present paper, we argue that a principled approach to comparing embeddings on the analogy task across languages consists in creating resources designed to be directly comparable. The most natural way of achieving this is by relying on manual translations, so as to retain a certain degree of control on the output quality and to produce resources that are maximally comparable. Given the weaknesses of GATS outlined by Gladkova et al. (2016), the more reasonable starting point for these translations would be the BATS dataset. These considerations effectively rule out the only similar dataset that we know of, by Ulčar et al. (2020), where analogies accept only one valid answer, as in GATS.

To that end, we introduce MATS, a Multilingual Analogy Test Set for six languages: Dutch, French, German, Italian, Mandarin, and Spanish, derived from the original BATS dataset of Glad-

[1]E.g., it is discussed by the Computerphile YouTube channel, cf. https://youtu.be/gQddtTdmG_8?t=662.

[2]To take an example, see the Winter 2017 NLP lectures at Stanford, https://youtu.be/ASn7ExxLZws?t=3257.

kova et al., spanning across 40 analogical relations equally partitioned between inflectional, derivational, lexicographic and encyclopedic. Using this new benchmark, we observe that different adaptations of the analogy task to mBERT contextual embeddings need not yield comparable results: Not only do we observe different performances when deriving static embeddings from contextual models and when using prompts, we also see that the exact wording of the prompt significantly impacts the model's behavior. We also share some anecdotal evidence questioning the validity of approaches to this task that assume there is a single gold answer—trained linguists attempting to solve this task often provide answers that do not match any of the expected targets, which further validates that single-target analogy benchmarks are ill-suited.

2 Related Works

Analogy, and specifically the offset approach of Mikolov et al. (2013b), has inspired the field at large (e.g., Roller et al., 2014; Bonami and Paperno, 2018; Ethayarajh, 2019; Chen et al., 2022). However, this approach has been criticized for methodological and ethical reasons (Bolukbasi et al., 2016; Linzen, 2016; Rogers et al., 2017; Schluter, 2018; Garg et al., 2018; Adewumi et al., 2022).

Two groups of related analogy datasets are often cited: those adapted from GATS (Google Analogy Test Set, Mikolov et al., 2013a) and those derived from BATS (Balanced Analogy Test Set, Gladkova et al., 2016). The latter distinguishes itself from the former on two major characteristics: First, it is designed for a balanced assessment of performances on the analogies and covers a larger collection of analogical relations; second, it admits multiple valid answers whenever relevant. These differences aim to mitigate some of the flaws Gladkova et al. (2016) perceived in GATS: The emphasis of this dataset on balance is intended to provide a more accurate picture of a model's capabilities when it comes to word analogy solving, and the inclusion of multiple answers aims to mitigate the impact of spelling variation and dataset limitations.

Datasets similar to BATS exist in Japanese and Icelandic (Karpinska et al., 2018; Friðriksdóttir et al., 2022), whereas GATS has been translated in Portuguese, Hindi, French, Polish, and Spanish (Hartmann et al., 2017; Grave et al., 2018; Cardellino, 2019). Other independently constructed datasets do exist (e.g., Venekoski and Vankka, 2017; Svoboda and Brychcín, 2018)—crucially, covering all languages of interest to this study: in Chinese (Jin and Wu, 2012; Chen et al., 2015; Li et al., 2018), Dutch (Garneau et al., 2021), English (Turney 2008; Mikolov et al. 2013b, a.o.), French (Grave et al., 2018), German (Köper et al., 2015), Italian (Berardi et al., 2015), and Spanish (Cardellino, 2019). On the other hand, these resources were created by different research groups and may contain items that are not easily comparable or of lesser quality.[3]

Similar to our approach, Grave et al. (2018) and Ulčar et al. (2020) both conduct multilingual comparisons of word embeddings on the analogy task, whereas Peng et al. (2022) study how analogies behave under cross-lingual mappings. All three works rely on GATS-style benchmarks (where only one valid target is admissible for each analogy relation); all are more limited in the scope of analogies they cover than BATS-style datasets; none study how contextual embeddings fit in this picture. This last point is partly due to the initial conception of the task for static models: Plenty of works discuss why static models develop linear analogies (Arora et al., 2016; Ethayarajh et al., 2019; Allen and Hospedales, 2019; Fournier and Dunbar, 2021)—similar evidence has yet to emerge for contextual models. As such, some studies delineate its relevance to static embeddings (e.g., Apidianaki, 2022), but it has been adapted to contextual models (Vulić et al., 2020; Ushio et al., 2021; Lenci et al., 2022).

3 The Multilingual Analogy Test Set

To study how analogy fares in a multilingual context, we introduce a Multilingual Analogy Test Set (MATS), adapted from BATS (Gladkova et al., 2016) for Dutch, French, German, Italian, Mandarin, and Spanish. This analogy benchmark is structured in two tiers: Individual sub-categories instantiating specific analogical relations (e.g., *country—capital*) are grouped into four general categories, namely **I**nflection, **D**erivation, **E**ncyclopedia, and **L**exicography. The former two correspond to morphological relations, such as the relation between two inflected forms of a word or the relation between a verb and the corresponding agent noun. The latter two are more closely aligned to commonsense reasoning and include relations such as synonymy or the relation between the name of a coun-

[3] E.g., the French dataset of Grave et al. (2018) mixes grammatical and social gender in masculine–feminine analogies.

try and that of its capital city. The original resource by Gladkova et al. (2016) emphasizes balance by ensuring that each of the four super-sections contains exactly 10 sub-sections, and that each of the 10 sub-sections contains exactly 50 instances of the same analogical relation; analogy quadruples are created by exhaustively iterating across pairs of instances. This totals to 98,000 distinct analogy quadruplets to test models on, around five times as many items as what is mentioned in Ulčar et al. (2020), and mitigates concerns of class imbalance.

Direct translations from the original BATS were taken as starting points before performing language-specific adaptations (cf. infra); we refer the reader to Gladkova et al. (2016) for supplementary details. In all languages, unidiomatic direct translations and analogically invalid pairs were removed. Multi-word expressions (MWE) were also removed,[4] before padding all categories except E03 to 50 pairs following the relation of each category. An overview of the outcome with examples and figures can be found in Table 1. We break down the choices per language in the following paragraphs.

Dutch The encyclopedic section E03 was localized using Dutch *provincies* and their capital cities.

French The inflectional section I03 was replaced with gender inflection of adjectives since comparatives are periphrastic constructions (e.g., *jolie* 'cute', *plus jolie* 'cuter'). The derivational section D01 was replaced with denominal adjectives using the suffix *-el*, as the formation of privatives using suffixes is not a productive morphological operation. The encyclopedic section E03 was localized using a random selection of 50 French *départements* and their capital cities, barring those that would be tokenized as MWE.

German The encyclopedic section E03 was localized with German *Länder* and their capital cities.

Italian The inflectional section I03 was replaced with gender inflection of adjectives, since Italian comparatives are periphrastic constructions (e.g., *bella* 'cute', *più bella* 'cuter'). The derivational section D01 was replaced with noun diminutives using the suffixes *-ino*, *-ina*, for the same reason as in French. The encyclopedic section E03 was localized using Italian *regioni* and their capital cities.

[4]Note this is a departure from BATS. This is for practical purposes, as we are also testing on static embeddings.

Mandarin Given the typological differences with English, we removed the whole section concerning inflectional morphology and completely reshaped the one on derivational morphology. In particular, given that derivation by means of affixes is a very productive process (Packard, 2000), we selected eight affixes, namely -度 '-ness/-ity', -化 '-ize', -性 '-ness/-ity', -学 '-ology', -主义 '-ism', -儿 'prosodic suffix', -机 'instrument', 小- 'diminutive prefix/small/young', and created corresponding categories. We set the focus of D09 on agent formation from verbs, much like D08 in all other languages, whereas for D10 we took inspiration from Li et al. (2018) focusing on reduplication of monosyllabic verbs having 'a bit' as semantic nuance. In the lexicographic category, we exploited elastic words (Guo, 1938; Duanmu, 2007) to build L08. We filled it using the list of elastic words in the Appendix of Dong (2015), focusing only on free monomorphemic adjectives and their corresponding long forms. The encyclopedic section E03 was localized using Chinese 省 and their capital cities. We incorporated the original E06 in D08 and replaced it with a category on nouns and their respective classifiers, disregarding the general classifier 个 that is not semantically informative.

Spanish The inflectional section I03 was replaced with gender inflection of adjectives since Spanish comparatives are periphrastic constructions (e.g., *linda* 'cute', *más linda* 'cuter'). The derivational section D01 was replaced with noun diminutives using the suffixes *-ito*, *-ita*, for the same reasons as in French and Italian. The encyclopedic section E03 was localized using Spanish *comunidades autónomas* and their capital cities.

4 Setting Baseline Expectations

We first focus on establishing the difficulty of our analogy benchmark, and how it compares to the English BATS. We provide a human baseline and static embedding scores on MATS.

Human Performance One aspect rarely addressed in analogy benchmarks is that of how consensual and accurate they are. Yet, some analogy relations are fundamentally debatable: For instance, whether "*tonne*" is to "*kilogram*" as "*flower*" is to "*petal*" depends on one's exact definition of a meronymic relation.[5] As such, the assumptions or intuitions of a given resource's designer may or

[5]These pairs are both in the L06 subcategory of BATS.

	de	es	fr	it	nl	zh
I01	*Tag : Tage*	*día : dias*	*jour : jours*	*dio : dèi*	*rol : rollen*	✗
I02	*Rat : Räte*	*voz : voces*	*bail : baux*	*base : basi*	*vlo : vlooien*	✗
I03	*süß : süßer*	*barato : barata*	*chanceux : chanceuse*	*colto : colta*	*oud : ouder*	✗
I04	*rein : reinste*	*feo : feísimo*	*drôle : drôlissime* (33)	*duro : durissimo*	*rijk : rijkst*	✗
I05	*hören : hört*	*crear : crea*	*dire : dit*	*godere : gode*	*vraagen : vraagt*	✗
I06	*teilnehmen : teilnehmend*	*creer : creyendo*	*gérer : gérant*	*gestire : gestendo*	*leren : lerend*	✗
I07	*sehen : gesehen*	*decir : dicho*	*croire : cru*	*perdere : perso*	*hoor : gehoord*	✗
I08	*glaubend : glaubt*	*girando : gira*	*lisant : lit*	*succedendo : succede*	*gaand : gaat*	✗
I09	*fragend : gefragt*	*uniendo : unido*	*ratant : raté*	*capendo : capito*	*vragend : gevraagd*	✗
I10	*wird : geworden*	*ejecuta : ejecutado*	*suit : suivi*	*sente : sentito*	*volgt : gevolgd*	✗
D01	*Arm : armlos*	*cabeza : cabecita*	*culture : culturel*	*stella : stellina*	*ego : egoloos*	强: 强度
D02	*fähig : unfähig*	*editado : inédito*	*pair : impair*	*certo : incerto*	*zeker : onzeker*	国际: 国际化
D03	*Kind : kindlich*	*real : realmente*	*fort : fortement*	*ampio : ampiamente*	*feest : feestelijk*	重要: 重要性
D04	*mäßig : übermäßig*	*poblado : sobrepoblado*	*aigu : suraigu*	*umano : sovrumano*	*vol : overvol*	语言: 语言学
D05	*fest : Festigkeit*	*fijo : fijeza*	*fou : folie*	*raro : rarità*	*vast : vastheid*	自由: 自由主义
D06	*geben : wiedergeben*	*mandar : remandar*	*lire : relire*	*spedire : rispedire*	*bouwen : herbouwen*	虫: 虫儿
D07	*haften : haftbar*	*evitar : evitable*	*jeter : jetable*	*vivere : vivibile*	*eeten : eetbaar*	打火: 打火机
D08	*tun : Täter*	*diseñar : diseñador*	*tuer : tueur*	*gestire : gestore*	*boksen : bokser*	孩子: 小孩子
D09	*reduzieren : Reduktion*	*acusar : acusación*	*priver : privation*	*mutare : mutazione*	*inspireren : inspiratie*	开发: 开发员
D10	*erklären : Erklärung*	*elevar : elevamiento*	*licencier : licenciement*	*pagare : pagamento*	*verklaren : verklaring*	想: 想想
L01	*Kuh : Wirbeltier/...*	*ganso : pájaro/...*	*caille : vertébré/...*	*ape : insetto/...*	*coyote : carnivoor/...*	猫头鹰: 鸟/...
L02	*Foto : Bild/...*	*sofá : mueble/...*	*bureau : objet/...*	*pompelmo : frutto/...*	*jas : eenheid/...*	架: 家具/...
L03	*Boot : Post/...*	*color : blanco/...*	*mois : décembre/...*	*canzone : inno/...*	*tasse : gral/...*	甜点: 蛋糕/...
L04	*Bart : Haar*	*agua : oxígeno/...*	*océan : eau*	*neve : acqua/...*	*staal : ijzer/...*	旗: 纸/...
L05	*Kalb : Vieh/...*	*cantante : coro/...*	*juré : jury*	*pecora : gregge*	*kal : vee/...*	鹅: 群
L06	*Byte : Bit*	*guitarra : cuerda/...*	*film : épisode/...*	*corpo : petto/...*	*euro : cent*	门: 铰链/...
L07	*ängstlich : entsetzt/...*	*amar : adorar/...*	*poney : cheval*	*triste : depresso/...*	*aap : gorilla*	湿: 浸泡/...
L08	*Fahrrad : Rad*	*madre : mamá*	*marché : bazar*	*roccia : sasso*	*vader : papa*	勇: 勇敢
L09	*heiß : frostig/...*	*claro : oscuro*	*sec : humide/...*	*sano : pazzo/...*	*jong : gaga/...*	甜: 酸/...
L10	*tot : lebendig*	*sucio : limpio*	*chute : montée*	*dopo : prima*	*west : oost*	内: 外
E01	*Lima : Peru*	*Bagdad : Irak*	*Damas : Syrie*	*Kiev : Ucraina*	*Zagreb : Kroatië*	安曼: 约旦
E02	*Iran : Persisch*	*Camboya : jemer*	*Égypte : arabe*	*Marocco : berbero/...*	*Cuba : Spaans*	伯利兹: 英语
E03	*München : Bayern* (13)	*Barcelona : Cataluña* (11)	*Nîmes : Gard* (50)	*Roma : Lazio* (17)	*Maastricht : Limburg* (10)	西安: 陕西 (27)
E04	*Marx : Deutsch*	*Homero : griego*	*Tolstoi : russe*	*Pascal : francese*	*Hegel : Duits*	孟子: 中国
E05	*Dante : Dichter*	*Depp : actor/...*	*Lincoln : président*	*Hawking : fisico/...*	*Locke : filosoof*	孔子: 哲学家
E06	*Ente : Küken*	*cigüeña : cigoñino*	*daim : faon*	*ape : larva*	*eend : eendje/...*	筷子: 双/...
E07	*Kuh : muhen*	*lobo : aúlla*	*hyène : rire*	*cane : abbaiare*	*ezel : balken/...*	猫: 喵/...
E08	*Wal : Meer/...*	*castor : río*	*bovin : étable*	*corvo : nido/...*	*beer : kooi/...*	狐狸: 洞穴
E09	*Kirsch : rot/...*	*peonía : roja/...*	*sel : blanc*	*tè : nero/...*	*bloed : rood*	蚂蚁: 黑色/...
E10	*Stier : Kuh*	*niño : niña*	*roi : reine*	*leone : leonessa*	*opa : oma*	老公: 老婆
Tot	1,963	1,961	1,983	1,967	1,960	1,477

Table 1: MATS: examples per subcategory. All subcategories contain 50 pairs, except if specified in (parentheses).

may not match with that of the community in general. Rare words may also factor in performances and dialectal variation can entail differences in spelling or vocabulary. Lastly, translation-based resources like ours may contain ambiguous cues and unknown cultural references.

So as to derive a human-level performance point of reference, for each language, we ask two trained linguists to manually solve 3 analogy items per subcategory, as well as two non-linguists for English[6] (cf. Appendix A). Annotators need not speak the same dialect, nor the dialect of the translators. While this may impact the reliability of the annotations, we choose to do so for two reasons. Firstly, the multiplicity of valid targets in the original BATS dataset was intended as a means to mitigate existing variations in the language at hand. Secondly, embeddings trained on large crawled corpora of internet texts will often span multiple dialects, and therefore factoring in linguistic variation provides a more principled point of comparison.

Annotators are provided with three of the four terms and ask them to propose a valid fourth term. We then measure (i) their **accuracy** on the task (i.e., the proportion of analogy items that were solved by the annotators with a valid fourth term in MATS) and (ii) their **agreement** rate (i.e., the proportion of analogy items where the two annotators produced the same answer).

Results in Table 2 show three global trends: (i) mistakes are made on almost all categories, (ii) linguistic training does help, and (iii) annotators' responses do not match 24%–46% of the time. Though these agreement scores may seem low, one ought to expect some variation across speakers in their ability to solve analogies—in part due to their familiarity with lexical semantics, in part due to dialectal variations between annotators, and in

[6]Results on English throughout this paper correspond to scores on Gladkova et al.'s BATS.

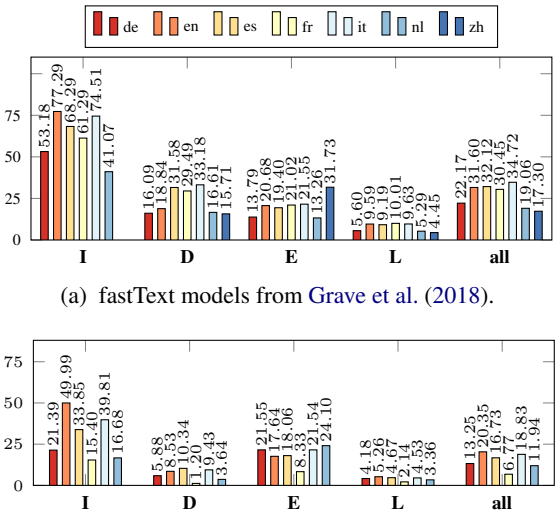
		Avg. accuracy					Agreement				
		I	D	E	L	all	I	D	E	L	all
en	ℓ	1.00	0.97	0.72	0.63	0.83	1.00	0.87	0.57	0.23	0.67
	$\neg\ell$	0.93	0.77	0.55	0.43	0.68	0.87	0.60	0.55	0.21	0.56
de		0.93	0.78	0.62	0.50	0.71	0.85	0.58	0.43	0.28	0.54
es		0.83	0.83	0.77	0.56	0.75	0.77	0.77	0.54	0.32	0.60
fr		0.88	0.97	0.70	0.48	0.76	0.83	0.93	0.52	0.30	0.65
it		0.97	0.93	0.75	0.57	0.80	0.93	0.86	0.81	0.42	0.76
nl		0.93	0.78	0.67	0.37	0.69	0.98	0.80	0.61	0.18	0.64
zh		—	0.85	0.62	0.35	0.61	—	0.83	0.57	0.43	0.61
all		0.92	0.86	0.67	0.49	—	0.89	0.78	0.51	0.30	—

Table 2: Manual annotations of MATS/BATS samples. $\ell/\neg\ell$: higher education in/not in linguistics.

part due to actual cases of linguistic ambiguity. In particular, we remark that both E and L include analogies that are less straightforward to solve for a human as compared with I and D, and some subcategories leave room for different interpretations due to their open-ended nature as described earlier. This is reflected in the overall lower accuracy and agreement scores for these two categories. In fact, annotators that indicate having looked up some of the analogy terms only report so for E and L. Crucially, performances on L are systematically the lowest, suggesting that this category is less in line with human reasoning.[7]

Static Embeddings Performance We now turn to static embeddings, which have been traditionally the target of analogy benchmarks. We consider two sets of available pre-trained static embeddings: the fastText models of Grave et al. (2018),[8] and the CoNLL-2017 Shared Task word2vec models (Zeman et al., 2017);[9] we set aside the CoNLL-2017 Chinese embeddings, as they correspond to traditional characters, whereas our resource is written in simplified characters.

We compute results on MATS, using the offset



(a) fastText models from Grave et al. (2018).

(b) word2vec models from Zeman et al. (2017).

Figure 1: Static models performance (`3CosAdd`, Equation (1)).

method of Mikolov et al. (2013b), a.k.a. `3CosAdd`:

$$\mathbf{b}_2^* = \underset{\mathbf{w}}{\operatorname{argmax}} \cos\left(\mathbf{w}, \mathbf{b}_1 + \mathbf{a}_2 + -\mathbf{a}_1\right) \quad (1)$$

This method consists in predicting as a target $\mathbf{b}_2^*$ the word w whose embedding $\mathbf{w}$ is the most codirectional to the offset-based approximation $\mathbf{b}_1 + \mathbf{a}_2 - \mathbf{a}_1$. The starting point of this approach is the assumption that for any two pairs of words instantiating the same semantic relation a_1, a_2 and b_1, b_2, their corresponding embeddings should be related by means of a stable offset. In other words, we assume that there exists a vector $\mathbf{x}$ such that

[7] It is also worth discussing the gap between English linguists and other languages: Beyond the variance that one expects given the very small sample size that was manually annotated, our English linguist annotators both use similar orthographic conventions as the original BATS resource; both also report a more extensive use of online search tools in case of doubts than annotators of other languages. Similar favorable conditions were never met for other languages. In short, the lower performances we observe for our resources should not be entirely imputed to them being translations.

[8] These cover 157 languages, including the seven of the present study. Note that their Chinese model corresponds to a mixture of traditional and simplified characters.

[9] Available at `http://vectors.nlpl.eu/repository/`.

$\mathbf{a}_1 + \mathbf{x} = \mathbf{a}_2$ and $\mathbf{b}_1 + \mathbf{x} = \mathbf{b}_2$, or equivalently $\mathbf{a}_2 - \mathbf{a}_1 = \mathbf{b}_2 - \mathbf{b}_1$, which we can reformulate to solve for $\mathbf{b}_2$ as $\mathbf{b}_2 = \mathbf{b}_1 + \mathbf{a}_2 - \mathbf{a}_1$. This method can therefore be seen as a direct assessment of whether analogical relations are encoded as stable offsets in the embedding space. In this work, we specifically rely on the `vecto` library implementation of `3CosAdd`.[10]

Results in Figure 1 show that fastText models perform better than CoNLL-2017 word2vec models, confirming the known trend (e.g., Bojanowski et al., 2017; Lenci et al., 2022). The noteworthy low performances on the L category across the board can be imputed to its lesser quality. In particular, fastText models score much higher for I and D, the two categories with morphological relations, likely thanks to their learning of character n-gram representations rather than word type representations—which makes fastText models overall more in line with manual annotations.

Beyond these general observations, language also impacts the scores we observe. For instance, the high scores observed for English word2vec on the I category are never attested for word2vec models in other languages—which can be pinned on the rather simplistic inflectional system in English. Both Dutch models along with the CoNLL-2017 French model perform surprisingly poorly. In the case of Dutch, this is likely due to training data limitations: Zeman et al. (2017) report training Dutch models on fewer than 3B words, whereas all other languages were trained on over 5B words.

Discussion The experiments conducted in Section 4 have helped us establish baseline expectations. Much of what we observe echoes previous findings: The improvement of fastText models on I and D analogy items was already documented in Bojanowski et al. (2017), and Levy and Goldberg (2014) or Gladkova et al. (2016) already highlighted lower performances on E and L analogies.

What is novel beyond these replicated findings is the observation that humans also struggle with E and L analogies. This can account in part for the lower performances observed for these categories. This also suggests that more lenient benchmarks like BATS, which allow multiple valid answers, are preferable to stricter ones, such as GATS.

[10] https://vecto.space/

	Sents	Tokens	Bytes	Types
de	300M	4.472B	28.448B	1.042M
en	300M	6.698B	35.396B	0.502M
es	300M	8.294B	46.133B	0.702M
fr	300M	6.058B	33.114B	0.581M
it	300M	7.266B	41.666B	0.631M
nl	300M	4.269B	24.320B	0.678M
zh	300M	15.594B	92.836B	1.531M

Table 3: Oscar corpora statistics. The last column tallies unique word types occurring at least 50 times.

5 Analogies and Contextual Embeddings

We now turn to benchmarking a contextual architecture, viz. uncased mBERT (Devlin et al., 2019). By definition, such architecture computes contextual representations of words: Unlike static embeddings, contextual embeddings vary depending on the entire input sequence. The default use-case intended for these models pertains to token-level semantics—whereas analogy benchmarks evaluate word-type-level semantics. One word may have different meanings depending on context—depending on which context we use, results on the task may vary drastically. This complicates the use of these representations for the analogy task, by introducing the need of deriving some form of type-level judgment from token-level representations.

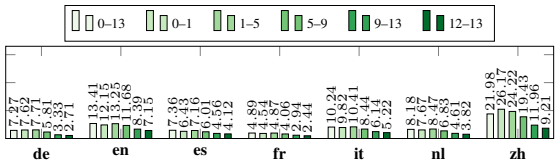
Static Representations from mBERT One possible approach to testing a contextual model on the analogy task consists in deriving word type representations from mBERT, and proceeding as one would with static embeddings. To determine which word types we need vectors for, we construct reference corpora of 300M sentences per language sampled from Oscar (Ortiz Suárez et al., 2019), and retrieve all word types with at least 50 occurrences.[11] All corpora were case-folded and tokenized using `spaCy`.[12] For Mandarin, we normalized all characters to their simplified form using `OpenCC`.[13] Corpora statistics are shown in Table 3.

We experiment with layer pooling and two different means of deriving static word-type vectors. **Singleton** embeddings are derived by embedding

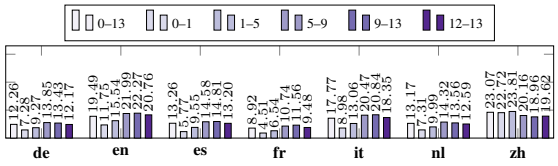
[11] This would correspond to a reasonable frequency filtering with word2vec embeddings, and matches what we used in supplementary experiments in Appendix C.

[12] https://spacy.io/

[13] https://pypi.org/project/OpenCC/

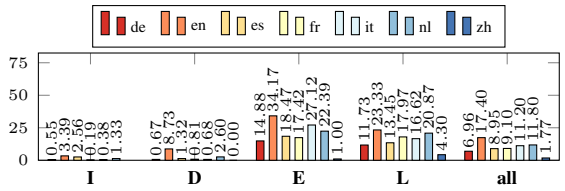


(a) Singleton embeddings.

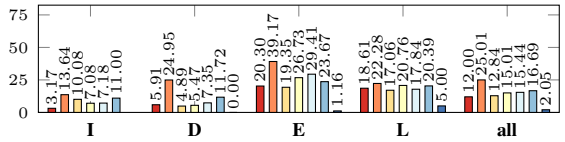


(b) Context-sample embeddings.

Figure 2: Static mBERT: overall results (3CosAdd).



(a) Unquoted templates.



(b) Quoted templates.

Figure 3: mBERT prompt-based performance.

word types as if they were simple sentences comprised of a single word and control tokens ([CLS] and [SEP]); we then sum across the whole sequence, and average over the layer representations of interest. For **context-sample** embeddings, we retrieve the first 10 contexts of occurrence of every word type[14] to compute the average embedding of that word type. In both cases, we draw representations from layers 0–1 (input embeddings), 12–13 (output vectors), 0–13 (all layers), 1–5, 5–9, and 9–13.

Overall accuracy results are displayed in Figure 2; results per category are available in Appendix B, Figure 7. Context-sample embeddings almost systematically outperform or equal the singleton approach for all layer groups and languages. Mandarin performs surprisingly well, and scores for all languages on the L category are extremely poor. With singleton embeddings, lower layers tend to perform better, which matches with previous studies (Vulić et al., 2020; Lenci et al., 2022), but performances for Mandarin are better when considering the embedding layer, whereas all other languages benefit most from pooling across the first four Transformer layers. On the other hand, European-language context-sample embeddings yield their highest performances with middle or top layers. We suspect that Mandarin has a very regular segmentation for D items, whereas Latin-alphabet languages may have different segmentations for otherwise regular suffixal construction, and therefore require some computation in order to properly reconstruct formal regularities. Scores per category provided in Figure 7, Appendix B confirm that much (almost all) of the performance attested for Mandarin is indeed driven by the D category.

Prompt-based Approaches Contextualized embeddings can also be tested by converting the task to a prompt format. We draw inspiration from the methodology of Ushio et al. (2021), but frame our analogies as an unmasking task. We fill a three-slot template $\mathcal{T}$ that contains a mask with three given analogy cues a_1, b_1, and a_2, and perform unmasking given the resulting sequence $\mathcal{T}(a_1, b_1, a_2)$. We measure a model's zero-shot accuracy by considering whether the unmasked word-pieces match with any of the listed valid targets' word-pieces.

All relevant templates are listed in Table 4. All templates were formulated by native speakers. In the case of targets split across multiple word-pieces, we include one mask token per word-piece; as such prompt scores are *stricto sensu* upper bounds.

Given the relative novelty of prompt-based approaches, we explore whether results are reliable across small changes of the prompts, such as the presence of quotation marks around analogy terms. Results in Figure 3 show that, besides English, performances are often lower than what we observed previously, and especially low on the I category. Prompts only outperform static vectors on the L category, which we established to be less reliable. Using quotes alleviates this trend, with a more pronounced effect on I and D. The higher English BATS scores are likely due to the large proportion of English training samples in mBERT.

We also test how behavior changes across semantically equivalent templates, using four alternative German templates, along with the effects of enquoting analogy terms. These templates are listed in Table 5. Results are displayed in Figure 4; the alternative template $\mathcal{T}_4$ corresponds to the default

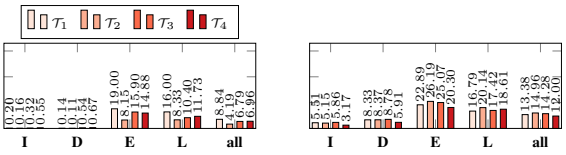
[14] We choose 10 contexts in order to strike a reasonable balance between diversity of contexts and computational costs.

	Unquoted	Quoted
de	a_1 verhält sich zu b_1 wie a_2 zu [MASK].	"a_1" verhält sich zu "b_1" wie "a_2" zu "[MASK]".
es	a_1 es a b_1 como a_2 es a [MASK].	"a_1" es a "b_1" como "a_2" es a "[MASK]".
fr	a_1 est à b_1 ce que a_2 est à [MASK].	"a_1" est à "b_1" ce que "a_2" est à "[MASK]".
it	a_1 sta a b_1 come a_2 sta a [MASK].	"a_1" sta a "b_1" come "a_2" sta a "[MASK]".
nl	a_1 staat tot b_1 zoals a_2 staat tot [MASK].	"a_1" staat tot "b_1" zoals "a_2" staat tot "[MASK]".
zh	a_1与b_1的关系就像a_2与[MASK]的关系。	「a_1」与「b_1」的关系就像「a_2」与「[MASK]」的关系。

Table 4: Templates for prompt-based approach.

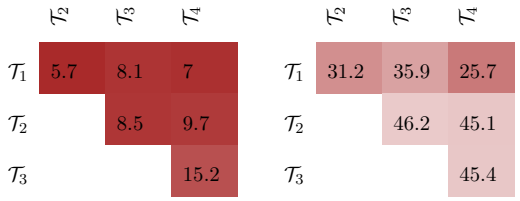
	Unquoted	Quoted
$\mathcal{T}_1$	de a_1 ist für b_1 was a_2 für [MASK] ist.	"a_1" ist für "b_1" was "a_2" für "[MASK]" ist.
$\mathcal{T}_2$	a_1 ist so zu b_1 wie a_2 zu [MASK] ist.	"a_1" ist so zu "b_1" wie "a_2" zu "[MASK]" ist.
$\mathcal{T}_3$	a_1 steht in Relation zu b_1 so wie a_2 zu [MASK].	"a_1" steht in Relation zu "b_1" so wie "a_2" zu "[MASK]".
$\mathcal{T}_4$	a_1 verhält sich zu b_1 wie a_2 zu [MASK].	"a_1" verhält sich zu "b_1" wie "a_2" zu "[MASK]".

Table 5: Alternative German templates.



(a) Unquoted templates. (b) Quoted templates.

Figure 4: Prompt-based performance of mBERT, using alternative German templates.



(a) Unquoted templates. (b) Quoted templates.

Figure 5: Prediction agreement (in %).

template for German in Figure 3. Quoted variants always outperform their unquoted counterparts; the model struggles most with the I and D categories. Yet, templates contrast starkly: E.g., by using the unquoted template $\mathcal{T}_1$ instead of $\mathcal{T}_2$, performance on E more than doubles, but this does not carry on with their quoted counterparts. In Figure 5, we tabulate how often predictions for the same analogy quadruple match across templates: Predictions of the mBERT uncased model tend to differ more often than they match, and this is much more pronounced with unquoted templates. In all, this model is sensitive to the exact wording of the prompt (cf. also Webson and Pavlick, 2022).

Discussion To sum up some key observations, we find mBERT ranks in between existing fastText and word2vec pre-trained embeddings. Results on the L category tend to be very low (except in the prompt-based approach). Scores for mBERT are highly dependent on methodology: Whether to include quotation marks in a prompt, or which layers static representations are derived from produce different effects across languages and categories.

All of this suggests that how to test contextual models like mBERT with analogies remains an open question. We observed different patterns across different languages and different methodologies. Some trends do emerge: For instance, static embeddings derived from mBERT do not appear to encode lexicographic and encyclopedic relations in any meaningful way, and Mandarin static mBERT embeddings are extremely apt at capturing derivational relationships, owing to their regular spelling. Likewise, recall that mBERT is not trained uniformly on all languages: This is most likely the reason why performance on English is higher. Prompt-based approaches, on the other hand, appear to capture E and L categories best, whereas I and D analogies are often poorly handled. This is the opposite of what we observed with human annotators in Section 4, which are more accurate on I and D rather than E and L items. Also worrying is the high volatility of the behavior: Prompt wording, or minor differences such as the presence or absence of quotes, can account for stark differences in the response patterns of mBERT.

For every methodological choice we explored—

which language and type of analogy to study, whether to use embeddings or prompts, how to derive the embeddings, or how to phrase the prompts—we observe distinct and often conflicting results. This is a direct consequence of the more complex architecture used in mBERT: The more varied means of probing and interacting with this model at our disposal also entail that we get a more diverse set of observations. As such, one can expect similar remarks to hold for other tasks. Establishing reasonable means of deciding which observations to select is both a captivating area for further inquiry and beyond the scope of this paper.

6 Conclusions

In this paper, we have presented a Multilingual Analogy Test Set, a resource five times larger than prior comparable datasets, with which we have looked at the analogy task in a multilingual context and studied how it fits in the modern NLP landscape. The dataset allows for a comparable multilingual evaluation of embedding models across a wide range of semantic analogy relations. Manual evaluation showed that the quality of MATS data in specific languages is comparable to the original English BATS. We saw that not all analogy types are equally straightforward not only to computational models but also to humans, and that behavior on the task depends on the language, the embedding model, and the methodology involved. This also entails that static model behavior is not a reliable indicator of what contextual models might yield.

We have been able to establish some trends across most of the methodological approaches we adopted here. In particular, from this work, we can outline three major conclusions. First, that not all categories are equally straightforward for humans (Section 4); this also explains why lower performances are attested on semantic analogies across most of our experiments. Second, that static models remain competitive with multilingual embedding models such as mBERT (Sections 4 and 5)—which replicates the conclusions of Lenci et al. (2022). Third, that equally valid prompts can yield vastly differing results (Section 5)—or more broadly, that different methodologies for adapting the analogy task to contextual embeddings can yield conflicting results. These conclusions also entail some practical guidelines for future work. In particular, there is a need to factor in human uncertainty as to what the correct target is; moreover, when adopting a prompt-based approach, testing a diverse array of prompts is necessary to properly establish how volatile a model's behavior is and how much variance in performance we should expect.

As such, a number of key challenges remain in the field of analogy solving, such as devising benchmarks that more closely match human intuitions or providing an explanatory framework for the discrepancies observed across prompts and methodologies. There are other aspects we have left open, such as whether the analogy task is suitable for lexical semantic evaluation (cf. Appendix C). We look forward to conducting future work in these directions, as well as expanding our observations to other architectures and methodologies.

Acknowledgment

We thank Alex Boulton, Guanyi Chen, Maria Copot, Yupei Du, Hermes Martínez, Giada Palmieri, Hugo Quené, Alisa Rieger, Vincent Segonne, William Soto, Joost Zwarts, and others for their help with the present work.

This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113). We also thank the CSC-IT Center for Science Ltd., for computational resources. This work was also supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program: Idex *Lorraine Université d'Excellence* (reference: ANR-15-IDEX-0004). This project has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement № 860621. This work was also supported by the ROCKY project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 742204).

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. Word2vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Computer Science*, 12(1):134–141.

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference*

on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231. PMLR.

Marianna Apidianaki. 2022. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, pages 1–60.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and Andre Freitas. 2018. SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio, Rivista semestrale*, 2/2018:173–196.

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1236–1242. AAAI Press.

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.

Yan Dong. 2015. *The prosody and morphology of elastic words in Chinese: annotations and analyses*. Ph.D. thesis, University of Michigan.

San Duanmu. 2007. *The phonology of standard Chinese*. OUP Oxford.

Kawin Ethayarajh. 2019. Rotate king to get queen: Word relationships as orthogonal transformations in embedding space. 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 3503–3508, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Louis Fournier and Ewan Dunbar. 2021. Paraphrases do not explain word analogies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2129–2134, Online. Association for Computational Linguistics.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2021. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12884–12892.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of*

the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Shaoyu Guo. 1938. 中国语词之弹性作用(The function of elastic word length in Chinese). *Yen Ching Hsueh Pao*, 24:1–34.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Peng Jin and Yunfang Wu. 2012. SemEval-2012 task 4: Evaluating Chinese word similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 374–377, Montréal, Canada. Association for Computational Linguistics.

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.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45, London, UK. Association for Computational Linguistics.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Lang. Resour. Eval.*, 56(4):1269–1313.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014,*

the 25th International Conference on Computational Linguistics: Technical Papers, pages 1025–1036, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana. Association for Computational Linguistics.

Lukáš Svoboda and Tomáš Brychcín. 2018. New word analogy corpus for exploring embeddings of czech words. In *Computational Linguistics and Intelligent Text Processing*, pages 103–114, Cham. Springer International Publishing.

Peter D. Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *J. Artif. Int. Res.*, 33(1):615–655.

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.

Viljami Venekoski and Jouko Vankka. 2017. Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 231–236, Gothenburg, Sweden. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Manual Annotation Details

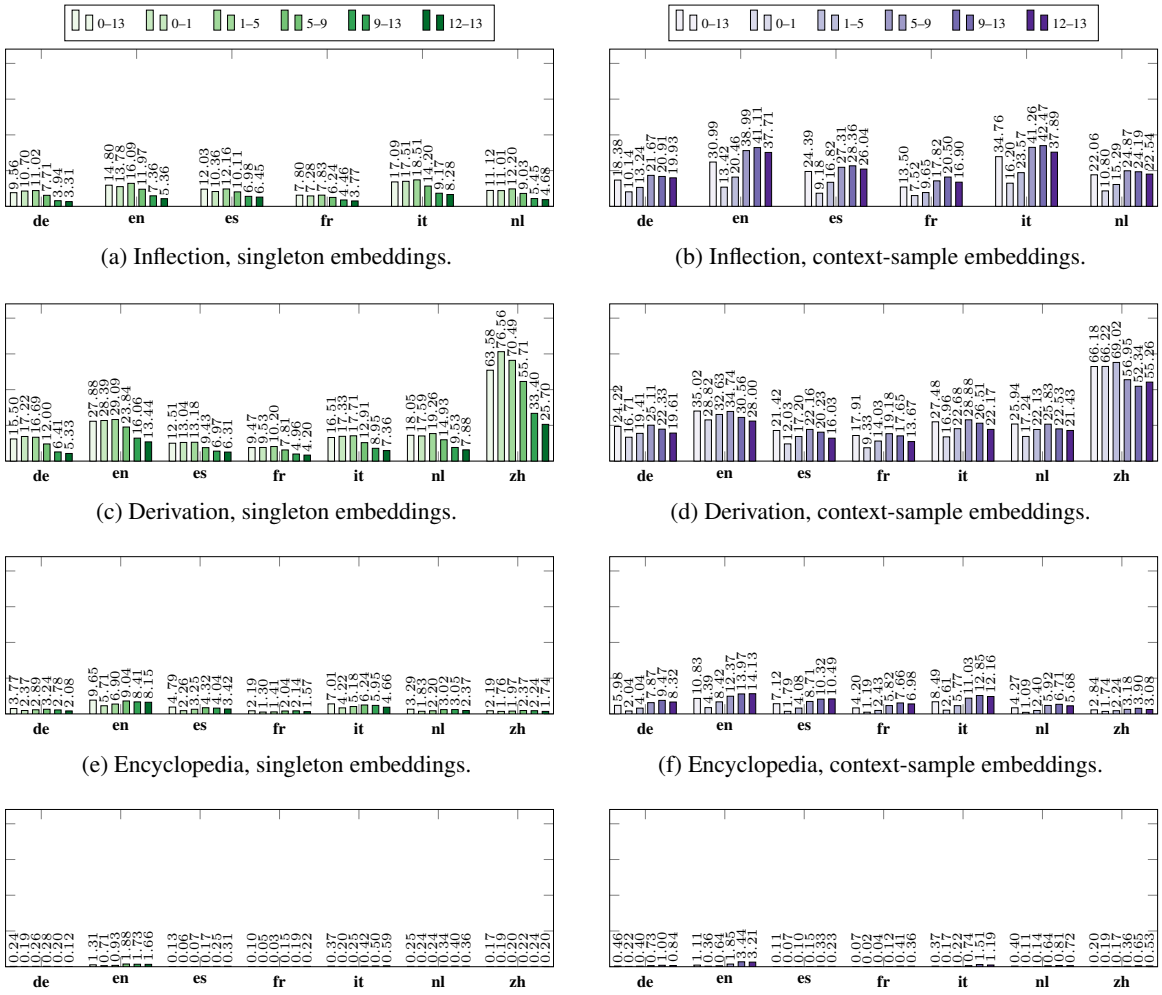
All annotators in Section 4 are volunteers and colleagues of the authors (or acquaintances, in the case of the two non-linguist English annotators), and are native speakers of the languages at hand. Provided instructions are shown in Figure 6.

> Each row is an incomplete analogy, please add your guess for the missing fourth term in a new column.
> For instance, given the three cues "king", "queen", "man", the fourth term ought to be "woman", since king is to queen as man is to woman.
>
> You can do multiple guesses, please put the one you're most confident about in first.
> For instance if you have a row where the three first columns are:
> `squirrel`, `squirrels`, `platypus`
> then fill the fourth column with
> *platypuses/platypi/platypodes*
> if you think "platypuses" is the most likely fourth term, but that "platypi" and "platypodes" are likely to be valid answers.
> All of your guesses should be single words.
>
> You are allowed to google things up if it helps: we are testing whether you can recover the relation, rather than whether you'd win at Jeopardy!.

Figure 6: Instruction provided to annotators.

B Detailed Results for Static mBERT

We provide per-category results for singleton and context-sample vectors on MATS in Figure 7. Key insights from Section 5 also hold for individual categories: Context-sample embeddings outperform



(a) Inflection, singleton embeddings.

(b) Inflection, context-sample embeddings.

(c) Derivation, singleton embeddings.

(d) Derivation, context-sample embeddings.

(e) Encyclopedia, singleton embeddings.

(f) Encyclopedia, context-sample embeddings.

(g) Lexicography, singleton embeddings.

(h) Lexicography, context-sample embeddings.

Figure 7: Static representations from mBERT: detailed results. All subplots share the same scale.

Param.	Values	Optimum on MATS					
		de	es	fr	it	nl	zh
window	$\{5, 10, 20\}$	20	20	20	5	20	20
neg. examples	$\{5, 10, 20\}$	10	20	5	20	20	20
shrink	$\{\top, \bot\}$	$\bot$	$\bot$	$\bot$	$\top$	$\bot$	$\top$
min freq.	$\{5, 50\}$	5	50	50	50	50	50
epochs	$\{1, 5\}$	5	5	5	5	5	5

Table 6: Hyperparameter search space.

singleton embeddings, and optimal layer groups vary across languages and categories.

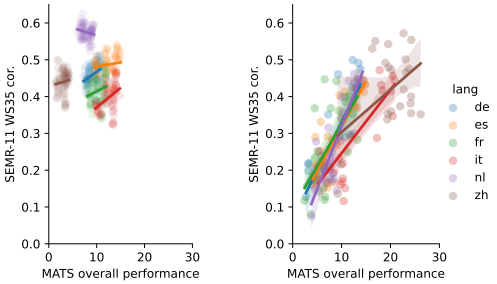
C Supplementary Experiment: Analogy vs. Semantic Similarity

An aspect we have not broached in the main body of this article is to what extent the analogy task is suitable to assess the semantic quality of the representations.

To answer this, we train 72 word2vec models per language with varying hyperparameters (cf. Table 6), on top of the static vectors derived from mBERT in Section 4 as well as similar static embeddings from the cased variant of mBERT, for a total of 24 mBERT-based static models per language.[15] Models were trained with `gensim` (Řehůřek and Sojka, 2010), using the reference corpus from Section 5. We then compare MATS overall accuracy scores to paired word cosine vs. human ratings correlation scores on the WS353 translations from Barzegar et al. (2018).

Results are displayed in Figure 8, and suggest that our static and contextual models behave differently. In the case of the former, the two benchmarks are not necessarily correlated (Table 7): While one can argue a trend exists for Italian and German, such a position is not supported for other languages.

[15] We ignore English to compare among translated benchmarks only.



(a) word2vec. (b) mBERT.

Figure 8: Behavior on MATS vs. on WS353.

		de	es	fr	it	nl	zh
w2v	cor.	0.42598	0.26613	0.29494	0.45498	-0.08502	0.15261
	p-val.	0.00019	0.02385	0.01190	0.00006	0.47765	0.20061
BERT	cor.	0.75217	0.78522	0.59913	0.64957	0.76174	0.29913
	p-val.	0.00002	0.00001	0.00198	0.00059	0.00002	0.15562

Table 7: Spearman correlation, WS353 vs. MATS.

As for mBERT, correlations appear to be reliable for all languages but Mandarin; note however that we have fewer observations than for word2vec. Furthermore, we notice little variation with word2vec, as highlighted by the clusters we get in Figure 8a.

In all, the behavior of earlier static models on lexical tasks such as similarity and analogy need not match with that of modern contextual embeddings. This also transpired in our earlier experiments: When comparing performances by category, the patterns we observe across categories seem quite specific to the architectures we test.

D Computational Costs

Throughout this paper, experiments involving mBERT have been performed using a single V100 GPU. This includes computing static embeddings and prompt-based scores. For the former, we observed variation across languages—e.g., Mandarin context-sample embeddings required over a day, but Dutch only took 4 hours. For the latter, processing one template took under 2 hours.

All other computations were run on clusters of 40 CPU cores. This includes training the word2vec models used in Appendix C, as well as running MATS and BATS evaluations for all static embeddings. Word2vec training scripts generally finished in under 4 hours. Evaluation runtimes on MATS and BATS depend on language, category, and vocabulary size, and range from under an hour to under a day per category (I, D, E, or L) and per model.

E Limitations

One limitation of our study is the inherent noisiness of the translations. Despite the language-specific adaptions, MATS is based on direct translations of BATS which was designed for English, and as such may not be entirely equivalent to a resource that has been specifically designed for the target languages. Gladkova et al. (2016) furthermore implemented datapoint selection criteria (such as a frequency-based filtering of target words) that we have not replicated in this work. Another element of quality control to address concerns the manual annotations in Section 4: Due to material limitations, annotations cover a very limited portion of the dataset and were conducted remotely.

Additionally, we only tested a few models in our study—word2vec and fastText for static embeddings and mBERT for contextual embeddings. This may not be representative of the full range of pre-trained language models, especially contextual ones. A similar point holds for the grid-search evaluation conducted in Appendix C. There are some word2vec hyperparameters we have not looked at and that could impact performances on both tasks: chief of which the dimension of the embeddings and the training corpus. More generally, expanding the number of models tested in future work could provide a more comprehensive understanding of the analogy task.

Another limitation is the lack of language diversity in our study. With the exception of Mandarin, all the languages we translated BATS into are Indo-European languages belonging to two sub-families (West Germanic or Romance languages).

Finally, the high computational power required to train the numerous word2vec models with varying hyperparameters in Section C (cf. Appendix D) both contributes to carbon emissions and limits the replicability of this work.