3D-EX: A Unified Dataset of Definitions and Dictionary Examples

Fatemah Almeman[*△] **Hadi Sheikhi**[°] **Luis Espinosa-Anke**[*◇]
[*]CardiffNLP, School of Computer Science and Informatics, Cardiff University, UK
[△] College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, KSA
[°] School of Computer Engineering, Iran University of Science and Technology, Iran
[◇]AMPLYFI, UK
{almemanf, espinosa-ankel}@cardiff.ac.uk
ha_sheikhi@comp.iust.ac.ir

Abstract

Definitions are a fundamental building block in lexicography, linguistics and computational semantics. In NLP, they have been used for retrofitting word embeddings or augmenting contextual representations in language models. However, lexical resources containing definitions exhibit a wide range of properties, which has implications in the behaviour of models trained and evaluated on them. In this paper, we introduce 3D-EX, a dataset that aims to fill this gap by combining well-known English resources into one centralized knowledge repository in the form of <term, definition, example> triples. 3D-EX is a unified evaluation framework with carefully pre-computed train/validation/test splits to prevent memorization. We report experimental results that suggest that this dataset could be effectively leveraged in downstream NLP tasks. Code and data are available at https://github.com/F-Almeman/3D-EX.

1 Introduction

Lexicographic definitions have played an important role in NLP. For example, definitions, and more specifically, term-hypernym pairs occurring in them, constitute a core component in applications such as taxonomy learning (Navigli et al., 2011; Velardi et al., 2013; Espinosa-Anke et al., 2016), knowledge base construction (Delli Bovi et al., 2015), or for augmenting language models (LMs) (Joshi et al., 2020; Chen et al., 2022). For this reason, numerous works have proposed methods to extract definitions from corpora (definition extraction, or DE) (Navigli and Velardi, 2010; Espinosa-Anke and Schockaert, 2018; Spala et al., 2020). However, DE, traditionally framed as a sentence classsification problem, plateaus quickly in terms of its applicability to real-world settings for a number of reasons, namely: (1) it is tied to a reference corpus; (2) it does not handle flexible contexts (e.g., definitional information appearing across several sentences); and (3) incorporating monolithic sentence-level definitional knowledge into LMs during pretraining is not straightforward. A complementary task to the above is definition modeling (DM), a promising direction both from resource creation and NLP standpoints. DM is the task of automatically generating human-readable lexicographic definitions or glosses given some input. From its inception, where Noraset et al. (2017) trained a bidirectional LSTM on $\langle t, d\rangle$ pairs, where t is an input term, and d is its corresponding definition, more recent contributions in this area have leveraged contextualized representations by augmenting t with some context c (Ni and Wang, 2017; Gadetsky et al., 2018; Ishiwatari et al., 2019; Reid et al., 2020; Bevilacqua et al., 2020).

A crucial prerequisite for enabling, among others, successful DM systems is having access to datasets that combine terms, definitions, and *good dictionary examples* (Kilgarriff et al., 2008; Kosem et al., 2019; Frankenberg-Garcia et al., 2019). In lexicographic resources, these good dictionary examples are written by professional lexicographers or domain experts, and often adhere to some style guidelines. This makes these sentences a valuable contextual resource for understanding the meaning of words, sometimes complementing knowledge gaps that may still exist even after reading a concept's definition.

DM is, arguably, one of the most recent direct NLP application of lexical resources. We therefore argue for the need of a centralized repository that could be used to train and test DM systems, explore out-of-domain generalization, and most importantly, act as a unified test bed for lexical semantics tasks. In this paper, we fill this gap by introducing 3D-EX, a dataset that unifies a diverse set of English dictionaries and encyclopedias. Our results suggest that, indeed, 3D-EX is a valuable

Proceedings of Recent Advances in Natural Language Processing, pages 69–79
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_008

resource for testing generative models in lexicographic contexts due to its varied sources, which makes it hard to memorize, and is also helpful for augmenting competitive baselines in downstream tasks.

2 Related work

Lexical resources have a long-standing tradition in lexical semantics (Camacho-Collados et al., 2018). Given the breadth of the area, we will review some of the most prominent existing resources, and then focus on how these resources have been leveraged in NLP tasks.

2.1 Lexical resources

Arguably, the best known lexical resource in NLP is WordNet (WN) (Miller, 1995), and as Hovy et al. (2013) described it, "the list papers using WN seems endless". Other resources which have complemented or augmented WN in the NLP space include knowledge bases such as Yago (Suchanek et al., 2008), DBPedia (Auer et al., 2007), BabelNet (Navigli and Ponzetto, 2012) or WikiData (Vrandečić and Krötzsch, 2014)[1]. Traditional dictionaries have also played an important role in NLP, we review these in Section 3, as they constitute the backbone of 3D-Ex.

2.2 Applications in NLP

Lexical resources in general, and dictionaries in particular, have played a critical role in recent years for improving (knowledge-rich and organic) NLP systems. For instance Faruqui et al. (2014) retrofitted word embeddings using semantic relations; Joshi et al. (2020) and Chen et al. (2022) used definitional information to augment pretrained LMs; and Delli Bovi et al. (2015), Espinosa-Anke et al. (2016) and Xu et al. (2022) used definitions for generating knowledge bases. In parallel, a generative avenue mostly revolving around DM has garnered substantial interest, where earlier works used LSTMs (Noraset et al., 2017; Gadetsky et al., 2018; Ishiwatari et al., 2019), and later contributions shifted to LMs (Bevilacqua et al., 2020; Huang et al., 2021; August et al., 2022). These works used DM models for downstream tasks like word sense disambiguation (WSD) (Navigli, 2009), word-in-context classification (Pilehvar and Camacho-Collados, 2019) or specificity-controlled glossary writing. Other works have explored complementary spaces, e.g., exemplification modeling (i.e., generating suitable dictionary examples given a word-definition pair) or full-fledged dictionary writing (Barba et al., 2021; de Schryver and Joffe, 2023; Sierra et al., 2023).

[1]Note that all these resources include definitions, unlike other resources designed for different purposes such as commonsense reasoning (e.g., ConceptNet (Speer et al., 2012)).

2.3 Datasets

Let us review the datasets we integrate into 3D-EX and how they have been applied either in lexicography or downstream NLP tasks.

WordNet: WN is an electronic lexical database for English that organises words in groups of synonyms called *synsets* (Miller, 1995; Fellbaum, 2013). Each synset is described by its definition, surface forms (lemmas), examples of usage (where available), and the relations between synsets, e.g., hypernymy (is-a), meronymy (is-part) or troponymy (manner-of). WN's primary use in NLP is as a sense inventory (Agirre and Edmonds, 2007; Zhang et al., 2022; Pu et al., 2023).

CHA: CHA (Chang and Chen, 2019) is an online dataset of words, definitions and dictionary examples from the Oxford Dictionary. It can be considered as a corpus of "traditional" dictionary definitions, and has been leveraged for DM by Bevilacqua et al. (2020) and for benchmarking the quality of WN's examples (Almeman and Espinosa-Anke, 2022).

Wikipedia: Wikipedia is an online encyclopedia that is created by various contributors on the web (Yano and Kang, 2016). In this work we used a dataset that is built by Ishiwatari et al. (2019) from Wikipedia and Wikidata and each entry consists of a phrase, description, and example. This dataset is used to evaluate DM approaches that combine distributional and lexical semantics using continuous latent variables (Reid et al., 2020).

Urban: Urban Dictionary is a crowd-sourced dictionary for terms that are not typically captured by traditional dictionaries (Wilson et al., 2020). In this work we used URBAN dataset that was created from Urban dictionary by Reid et al. (2020) as a corpus of uncommon and slang words.

Wiktionary: Wiktionary is a freely available web-based dictionary that provides detailed information on lexical entries such as definitions, examples of usage, pronunciation, translations, etc.

(Bajčetić and Declerck, 2022). It has been used as a resource for WSD (Meyer and Gurevych, 2011; Matuschek and Gurevych, 2013), especially for retrieving WSD examples which augment labeled data for rare senses (Blevins et al., 2021) and for non-English tasks (Henrich et al., 2012; Segonne et al., 2019).

Webster's Unabridged: Webster's Unabridged is a version of Webster's dictionary (Webster, 1900) served by the Project Gutenberg initiative (Various, 2009). It describes English words by providing definitions and notes (where needed).

Hei++: Hei++ is a dataset that associates human-made definitions with adjective-noun phrases. Since there is no publicly available dataset to evaluate the quality of definition generation models on free phrases, Hei++ is built by Bevilacqua et al. using the test split of the HeiPLAS dataset (Hartung, 2015).

MultiRD: The MultiRD dataset was created by (Zhang et al., 2019) to evaluate a multi-channel reverse dictionary model that has multiple predictors to predict attributes of target words from given input queries. This dataset uses the English dictionary definition dataset created by Hill et al. (2016) as the training set and three test sets: a *seen* definition set, an *unseen* definition set, and a description set that includes pairs of words and human-written descriptions. For each entry, it also includes morphemes, lexical names and sememes.

CODWOE: The CODWOE (Comparing Dictionaries and Word embeddings) SemEval 2022 shared task (Mickus et al., 2022) aimed to compare two types of semantic descriptions, namely dictionary glosses and word embedding representations. This task was applied to multiple languages, and one dataset per language was provided. Each dataset contains a list of examples and, subsequently, each example contains the following key fields: identifier (includes the word), gloss, and embedding-related information.

Sci-definition: Sci-definition is a dataset constructed for the task of generating definitions of scientific terms with controllable complexity (August et al., 2022). The definitions are drawn from MedQuAD (Abacha and Demner-Fushman, 2019) and Wikipedia Science Glossaries[2]. For each term, 10 journal abstracts are provided from S2ORC (Lo et al., 2020) to allow models to incorporate related scientific knowledge (Fan et al., 2019; Clark et al., 2018).

[2] https://en.wikipedia.org/wiki/Category:Glossaries_of_science.

3 Building 3D-EX: Data Cleaning

A prerequisite for unifying the above resources into 3D-EX, is to perform a number of preprocessing steps. This process includes: lower-casing; removing special tokens and any noisy characters such as the `tab` sign; removing entries where their definitions have more than 10% of non alphanumeric characters; removing entries that have null values either in words or definitions; removing entries where examples are the same as defined terms, and removing duplicate entries within each dataset or split.

3.1 Dataset-specific cleaning

While the above steps are applied to all datasets, each individual resource in 3D-EX undergoes a specific preprocessing set of steps:

Urban: since Urban dictionary is built by end-users who are not trained lexicographers, we found that it has number of noisy definitions (typically, too short, or containing a high proportion of emoticons, exclamation marks, and so forth). To handle them, we built a binary classifier based on RoBERTa-base (Liu et al., 2019) where 4,000 positive examples are randomly sampled from Wiktionary, CHA and WN, and 2,000 negative examples are randomly sampled from Urban. This classifier, which obtains almost perfect accuracy, is then applied to the entirety of the Urban dataset, leaving 3D-EX only with Urban entries that are similar to those in more traditional resources, both in content and, more importantly, in style. Table 1 lists examples of this filtering process, where we can see Urban-specific properties such as colloquialisms (phrasal verbs, personal pronouns, lack of punctuation marks or high proportion of slang/unknown words).

Wiktionary: Since some definitions in Wiktionary include the time where words were coined (e.g., "first attested in the late 16th century" or "from 16 c"), we deleted them using regular expressions.

MultiRD: we removed (again, using regular expressions) uninformative definitions such as "see synonyms at" and "often used in the plural".

Term	Definition	Example	F.
baby bentley	a way to describe a beat up old car you wish was a Bentley	Dave calls his beat-up Neon his baby Bentley	1
pang	pangers pingerz pang pangs pangs MDMA ecstasy	Hi Marissa, it's Frank Record calling. I'll be in the neighborhood later on, and I was wondering if maybe you wanted to get some pang pangs	1
suckafish	the correct term for one who you think is a sucker, loser, or anything else	Wow, that guy is being a total suckafish	1
farblegarb	a lot of random garbage	The signal was disrupted, producing a lot of farblegarb	0
citrixify	the process of modifying or altering a computer application for the purpose of publishing the application using Citrix Presentation Server	In order to properly publish that Java-based application, I had to citrixify it so it would run in a seamless window	0
axcellent	when something rocks and is excellent	Dude, that new haircut is axcellent	0

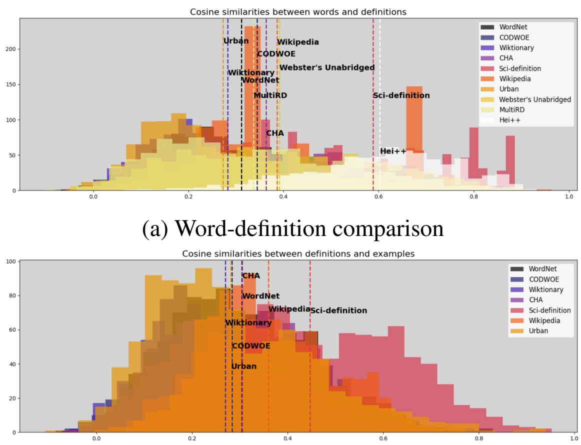
Table 1: Examples of Urban entries that were removed vs. retained (labels 1 vs. 0 in column **F.**).

Sci-definition: in order to construct the **Sci-definition** dataset as <`term, definition, example`> triples, we took the following steps: from each abstract, we extracted sentences that include the target term, which would act as examples. From these examples, we excluded sentences only containing lists of keywords (typically found in abstracts), and also any example with more than 10% non alphanumeric characters (similarly to our approach to cleaning definitions in Section 3).

3.2 Unification and splitting

Tables 2 and 3 show summary statistics for each dataset. It is desirable to keep a reference to the original source (dictionary or glossary) for each entry, however, we noticed that there are <`term, definition, example`> duplicates across datasets. This is why the final 3D-EX resource contains the SOURCE field as an array containing the sources where that entry was found. Furthermore, in terms of splitting 3D-EX for experimentation, it is well known that an issue in word/phrase classification datasets can occur due to a phenomenon known as "lexical memorization" (Levy et al., 2015), where supervised models tend to associate prototypical features to word types. This has been typically been addressed by releasing two splits, one random, and one known as "the lexical split", where all instances of a given term do not appear across splits (Vulić et al., 2017; Apidianaki and Soler, 2021; Espinosa-Anke et al., 2022). We follow this practice and release 3D-EX with a Random and a Lexical split. Tables 4 and 5 show examples of entries in 3D-EX and dataset statistics after unification in terms of unique instances across both splits, respectively.

Finally, to shed some light on how similarities are distributed across datasets, we investigate cosine similarities of their SBERT embeddings, and compute similarities between terms and definitions, and between definitions and examples (see Figure 1). An immediate finding by inspecting these similarities is that Hei++, a carefully curated dataset used to evaluate multiword DM systems, is the one showing the highest similarity between terms and definitions (Figure 1a), this is likely because, first, entries in Hei++ are rather specific, and do not include generic and frequently used terms. This, along with, also, a rather detailed definition, makes their similarity rather high. On the opposite end of the spectrum we unsurprisingly find Urban dictionary, although it remains for future work to explore whether Urban Dictionary's definitions are indeed dissimilar to their corresponding terms, or because they are so rare that their embeddings are of lower quality. Interestingly, we also find that Sci-definition also exhibits high similarity between terms and definitions. Concerning definitions and examples (Figure 1b), Sci-definition is again the one with the highest similarity scores, and interestingly, Wiktionary is the dictionary with the lowest aggregate similarity, which suggests that examples in Wiktionary could be purposefully written to cover different topics than their definitions. As with the case of Urban Dictionary, a careful semantic analysis of these dictionaries remains for future work.



(a) Word-definition comparison

(b) Definition-example comparison

Figure 1: Histograms with SBERT-based cosine similarities of the datasets in 3D-EX.

	orig. #entries	cl. #terms	cl. # <T,D>	cl. #<T,D,E>
WordNet	44,351	20,435	36,095	44,241
CHA	785,551	30,841	75,887	752,923
Wikipedia	988,690	162,809	167,569	960,097
Urban	507,638	119,016	145,574	145,896
Wiktionary	145,827	76,453	85,905	140,190
CODWOE	63,596	25,861	45,065	63,137
Sci-definition	8,263	5,281	6,251	166,660
Webster's Unabridged	159,123	89,234	143,782	-
MultiRD	901,200	50,460	671,505	-
Hei++	713	713	713	-
3D-EX		438,956	1,327,342	2,268,225

Table 2: Dataset statistics before (orig.) and after (cl.) preprocessing, and in terms of unique entries involving terms (**T**), definitions (**D**), examples (**E**). Aggregated statistics are provided between two sets, datasets with examples (top) and without (bottom). The last row is related to 3D-EX dataset.

	Term length			**Definition length**			**Example length**		
	min.	max.	avg.	min.	max.	avg.	min.	max.	avg.
WordNet	1	1	1	1	52	7.50	1	46	5.77
CHA	1	1	1	1	71	10.31	2	141	17.86
Wikipedia	1	16	1.84	1	32	6.012	2	40	18.70
Urban	1	31	1.47	1	32	10.01	2	42	11.45
Wiktionary	1	10	1.22	1	100	9.24	2	288	26.52
CODWOE	1	1	1	1	114	10.86	1	214	22.26
Sci-definition	1	11	1.70	2	94	18.49	1	726	25.72
Webster's Unabridged	1	3	1.00	1	90	9.19	-	-	-
MultiRD	1	1	1	1	144	11.72	-	-	-
Hei++	2	2	2	3	23	8.12	-	-	-

Table 3: Length statistics per dataset after cleaning.

4 Experiments and Results

In order to test the usefulness of 3D-EX, we perform an intrinsic set of experiments where we "stress test" the dataset for artifacts, indirect data leakage (near-synonyms), potential for memorization, etc. This, we argue, is an important step to guarantee 3D-EX can be used for testing lexical semantics models based on it.

4.1 Source classification

In the task of *source classification*, the goal is to, given a <term,definition> instance, predict its original source. We posit that this is an important experiment to determine which sources are more unique (i.e., easier to classify), and which seem to conflate different lexicographic features (e.g., writing style, coverage or any other artifact). To this end, we fine-tune `roberta-base` (Liu et al., 2019) for 3 epochs on the training set of 3D-EX. Note that this is a 9-way multilabel classification problem, since for a given <term,definition> tuple, there may be more than one associated source.

We report the results of this experiment in Table 6. We can see how the lexical split is substantially harder than the random split.

4.2 Reverse dictionary

Reverse dictionary (or concept finder) is a helpful application for copywriters, novelists, translators seeking to find words or ideas that might be "on the tip of their tongue" (Hill et al., 2016). It is also reflection of the interactions between a speaker and the mental lexicon (Zock, 2004; Zock et al., 2010). More relevant to NLP, however, reverse dictionary datasets can be seen as benchmarks for evaluating representation learning methods, as there are works that have used definitions as, e.g., the sole source for learning word embeddings (Bosc and Vincent, 2017) or for debiasing them (Kaneko and Bollegala, 2021).

This task is a ranking problem in which, given a definition, the task is to retrieve a ranked list of the most relevant words, and it has a long-standing tradition in computational semantics (Bila et al., 2004; Dutoit and Nugues, 2002; El-kahlout and Oflazer, 2004; Glassman et al., 1992; Thorat and Choudhari, 2016) . To establish a set of baseline results on this task, we report results from several embedding models on the random and lexical test sets. Note that while these baselines are unsupervised, we only report results on the test sets to accommodate future experiments by supervised systems. In terms of evaluation, we report *Mean Reciprocal Rank* (MRR), which rewards the position of the first correct result in a ranked list of outcomes:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where Q is a sample of experiment runs and $rank_i$

Term	Definition	Example	source
emergent	coming into existence	an emergent republic	WordNet
word	an (order; a request or instruction); an expression of will	he sent word that we should strike camp before winter	Wiktionary
central london	innermost part of london , england	westminster is an area of central london within the city of westminster , part of the west end , on the north bank of the river thames	Wikipedia
ejac-flashback	when a picture or video is familiar to you	dude I've just had a ejac-flashback that chick was last nights wank material	Urban
notice	a displayed sheet or placard giving news or information	look out for the notice of the samaritans information evening in the end of september	CHA
worship	to participate in religious ceremonies	we worship at the church down the road	CODWOE
accessory navicular bone	an accessory navicular bone is a small bone located in the middle of the foot	the accessory navicular bone is one of the most common accessory ossicles, which sometimes become symptomatic	Sci-definition
able	having sufficient power, strength, force, skill, means, or resources of any kind to accomplish the object	-	Webster's Unabridged
abbreviation	an abbreviation is a shorter way to write a word or phrase	-	MultiRD
skew picture	an inaccurate or partial representation of a situation	-	Hei++

Table 4: Examples of entries available in 3D-EX.

	Random split			Lexical split		
	train	**validation**	**test**	**train**	**validation**	**test**
WordNet	26,603	8,788	8,850	27,053	8,573	8,793
CHA	451,191	15,1338	50,394	452,321	157,847	143,949
Wiktionary	84,111	28,127	27,952	89,607	29,176	23,832
Wikipedia	575,554	197,697	186,846	505,964	240,781	213,379
Urban	87,429	29,142	29,325	91,239	29,783	24,881
CODWOE	37,774	12,755	12,608	39,737	12,609	13,166
Sci-definition	101,129	31,766	33,765	106,175	35,966	24,519
Webster's Unabridged	84,802	28,213	28,221	93,423	30,198	19,696
MultiRD	384,295	127,580	128,178	404,114	125,072	112,948
Hei++	426	152	135	428	143	142

Table 5: Breakdown of 3D-EX unique entries per split type (random and lexical) and per split. Note that unique entries consist of <term,def.,example,source> (first 6 rows) or <term,def.,source> (bottom 3 rows).

	Random Split			Lexical Split		
	prec.	rec.	f1	prec.	rec.	f1
WordNet	0.73	0.23	0.35	0.33	0.05	0.09
CHA	0.65	0.48	0.55	0.64	0.47	0.54
Wiktionary	0.80	0.53	0.64	0.65	0.33	0.44
Wikipedia	0.98	0.97	0.98	0.97	0.97	0.97
Urban	0.94	0.87	0.91	0.97	0.66	0.79
CODWOE	0.93	0.55	0.69	0.92	0.42	0.58
Sci-definition	0.99	0.99	0.99	0.99	0.99	0.99
Webster's Unabridged	0.82	0.70	0.76	0.75	0.63	0.68
MultiRD	0.89	0.90	0.89	0.84	0.91	0.88
Hei++	0	0	0	0	0	0
Average	0.77	0.62	0.68	0.71	0.54	0.60

Table 6: Results in the source classification experiment, reported both for the Random and Lexical splits of 3D-EX.

refers to the rank position of the *first* relevant outcome for the *i*th run. MRR is commonly used in Information Retrieval and Question Answering, but has also shown to be well suited for lexical semantics tasks such as collocation discovery (Wu et al., 2010; Rodríguez-Fernández et al., 2016).

We evaluate the performance of traditional sentence encoding SBERT (Reimers and Gurevych, 2019) models, namely `all-MiniLM-L6-v2`, `all-distilroberta-v1` and `all-mpnet-base-v2`. We also evaluate Instructor (Su et al., 2022), an instruction-based encoder that can generate text embeddings tailored to any task given the appropriate prompt. Instructor works by optionally providing the type of the target text (e.g., "a Wikipedia sentence") and the task (e.g., "document retrieval"), to ultimately build a prompt such as "Represent this Wikipedia sentence for

Model	Random	Lexical
all-distilroberta-v1	8.41	11.38
all-MiniLM-L6-v2	9.40	13.75
all-mpnet-base-v2	10.98	15.34

Table 7: Reverse Dictionary results of the SBERT models on the reverse dictionary task in the two 3D-EX test sets.

retrieving relevant documents". For our use case, we test three variants of Instructor for encoding both words and definitions: (1) no instruction; (2) providing a generic description of the target text (i.e., "the sentence" and "the word"); and (3) providing a domain-specific description of the target texts (i.e., "the dictionary definition" and "the dictionary entry").

We show the results of the SBERT models in Table 7, and the Instructor results in Table 8. We can see that even without any instruction prepended to the embedder, the Instructor model outperforms vanilla SBERT models, and that, interestingly, the best results overall in both splits (random and lexical) are obtained by providing a generic description of target words, and in the random split it is better to not include instructions for the definitions, while in the lexical split the best performing configuration involves providing detailed instructions for embedding the 3D-EX definitions.

As a final piece of analysis, we perform experiments on both test sets with the best performing model (based on the split type) to see which sources are harder to solve in the task of reverse dictionary. From Table 9, it can be seen that Wikipedia and Urban are the most challenging resources for this task, which could be attributed to either or both dataset size and large number of very similar definitions and terms, as opposed to for instance Hei++ or Sci-definition, which are meant to capture unique terms. These are, by nature, more unique when compared to the rest of the lexicon, an insight we revealed when exploring dataset-specifc similarities in Figure 1.

Random		word: no	word: gen.	word: dict.
definition	no	14.18	**14.71**	14.56
	gen.	13.64	14.07	14.06
	dict.	14.19	14.59	14.57
Lexical		**word: no**	**word: gen.**	**word: dict.**
definition	no	19.16	20.25	20.02
	gen.	18.70	20.04	19.86
	dict.	19.64	**20.82**	20.60

Table 8: MRR Results on Reverse Dictionary leveraging Instructor Embeddings when using no instruction (no), generic (gen.) or tailored to the task (dict.).

Dataset	Random	Lexical
WordNet	32.97	42.27
Wiktionary	50.65	53.05
Wikipedia	9.25	9.19
Urban	18.47	17.49
CODWOE	39.74	46.89
CHA	30.82	35.86
Sci-definition	82.38	82.53
Webster's Unabridged	30.53	34.11
MultiRD	16.69	27.41
Hei++	96.79	94.49

Table 9: Breakdown of the reverse dictionary results in terms of MRR for the two test sets (random and lexical) in 3D-EX.

5 Conclusions and future work

In this paper we have introduced 3D-EX, a dataset that unifies different encyclopedias and dictionaries into one single resource. We have conducted an in-depth analysis of the dataset across several splits (random vs lexical), as well as dictionary source classification and reverse dictionary experiments. Our results suggest that this dataset is both challenging for representation learning methods and promising as a resource for augmenting lexical semantics systems. It has also helped us unveil semantic properties in the different dictionaries and encyclopedias we have integrated into 3D-EX.

For the future, we would like to further explore the potential of 3D-EX for downstream NLP tasks, incorporating more resources, and exploring multilingual variants. An additional avenue would be to explore the interaction of unorthodox dictionaries like Urban with traditional lexicographic resources in the context of controlled technical/jargon DM. Finally, leveraging 3D-EX as a resource for pretraining LMs, similarly to the DictBERT approach (Chen et al., 2022), could help inform LMs with new, domain-specific and/or colloquial terms.

Ethics and Broader Impact Statement

This paper is concerned with the automatic building of a dataset by combining publicly available information in the web. As a result, there could be potential for the presence of incorrect or harmful information in this derived dataset, especially if crowdsourced; however, we encourage collaborative efforts from the community to help address these risks. Specifically, vulgar, colloquial, or potentially harmful information in Urban Dictionary, which the authors of this paper do not endorse.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1).

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Fatemah Almeman and Luis Espinosa-Anke. 2022. Putting wordnet's dictionary examples in the context of definition modelling: An empirical analysis. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 42–48.

Marianna Apidianaki and Aina Garí Soler. 2021. All dolphins are intelligent and some are friendly: Probing bert for nouns' semantic properties and their prototypicality. *arXiv preprint arXiv:2110.06376*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Lenka Bajčetić and Thierry Declerck. 2022. Using Wiktionary to create specialized lexical resources and datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3457–3460, Marseille, France. European Language Resources Association.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3779–3785. International Joint Conferences on Artificial Intelligence Organization.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Slaven Bila, Wataru Watanabe, Taiichi Hashimoto, Takenobu Tokunaga, and Hozumi Tanaka. 2004. Dictionary search based on the target word description.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. Fews: Large-scale, low-shot word sense disambiguation with the dictionary.

Tom Bosc and Pascal Vincent. 2017. Learning word embeddings from dictionary definitions only. In *Proceedings of the NIPS 2017 Workshop on Meta-Learning*.

Jose Camacho-Collados, Luis Espinosa Anke, and Mohammad Taher Pilehvar. 2018. The interplay between lexical resources and natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–23.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. 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 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Qianglong Chen, Feng-Lin Li, Guohai Xu, Ming Yan, Ji Zhang, and Yin Zhang. 2022. Dictbert: Dictionary description knowledge enhanced language model pre-training via contrastive learning. *arXiv preprint arXiv:2208.00635*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Dominique Dutoit and Pierre Nugues. 2002. A lexical database and an algorithm to find words from definitions.

Ilknur El-kahlout and Kemal Oflazer. 2004. Use of wordnet for retrieving words from their meanings.

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Luis Espinosa-Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. 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 378–385.

Luis Espinosa-Anke, Alexander Shvets, Alireza Mohammadshahi, James Henderson, and Leo Wanner. 2022. Multilingual extraction and categorization of lexical collocations with graph-aware transformers. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 89–100.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Christiane Fellbaum. 2013. Wordnet. In Carol Chapelle, editor, *The encyclopedia of applied linguistics*, pages 6739–6746. Blackwell Publishing Ltd.

Ana Frankenberg-Garcia, Robert Lew, Jonathan C Roberts, Geraint Paul Rees, and Nirwan Sharma. 2019. Developing a writing assistant to help eap writers with collocations in real time. *ReCALL*, 31(1):23–39.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

L Glassman, Dennis Grinberg, Cynthia S. Hibbard, and James C. Meehan. 1992. Hector: Connecting words with definitions.

Matthias Hartung. 2015. *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2012. Webcage: a web-harvested corpus annotated with germanet senses. pages 387–396.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. 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 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. *arXiv preprint arXiv:2101.09525*.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Barcelona, Spain. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Iztok Kosem, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, and Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: the case (s) of gdex. *International Journal of Lexicography*, 32(2):119–137.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2orc: The semantic scholar open research corpus.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

Christian M. Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *International Joint Conference on Natural Language Processing*.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1318–1327.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, volume 11, pages 1872–1877.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. pages 3259–3266.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Pu, Lin Yuan, Jiaxu Leng, Tao Wu, and Xinbo Gao. 2023. Lexical knowledge enhanced text matching via distilled word sense disambiguation. *Knowledge-Based Systems*, page 110282.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6331–6344.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sara Rodríguez-Fernández, Luis Espinosa Anke, Roberto Carlini, and Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–505.

Gilles-Maurice de Schryver and David Joffe. 2023. The end of lexicography, welcome to the machine: On how chatgpt can already take over all of the dictionary maker's tasks. In *20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics*.

Vincent Segonne, Marie Candito, and Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden. Association for Computational Linguistics.

Óscar García Sierra, Miguel Ortega-Martín, Alfonso Ardoiz, Juan Carlos Armenteros, Jorge Álvarez, and Adrián Alonso. 2023. Spanish built factual freectianary (spanish-bff): the first ia-generated free dictionary. *arXiv preprint arXiv:2302.12746*.

Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. Semeval-2020 task 6: Definition extraction from free text with the deft corpus. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 336–345.

Robyn Speer, Catherine Havasi, et al. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, volume 2012, pages 3679–86.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217.

Sushrut Thorat and Varad Choudhari. 2016. Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. In *Proceedings of COLING 2016, the 26th International Conference on*

Computational Linguistics: Technical Papers, pages 2797–2806, Osaka, Japan. The COLING 2016 Organizing Committee.

Various. 2009. *Webster's Unabridged Dictionary*. Project Gutenberg.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Noah Webster. 1900. *Webster's unabridged dictionary of the English language*. Kikwansha.

Steven R. Wilson, Walid Magdy, Barbara McGillivray, Venkata Rama Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang nlp applications. In *International Conference on Language Resources and Evaluation*.

J.C. Wu, Y.C. Chang, T. Mitamura, and J.S. Chang. 2010. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.

Hongyuan Xu, Yunong Chen, Zichen Liu, Yanlong Wen, and Xiaojie Yuan. 2022. Taxoprompt: A prompt-based generation method with taxonomic context for self-supervised taxonomy expansion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4432–4438. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tae Yano and Moonyoung Kang. 2016. Taking advantage of wikipedia in natural language processing.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2019. Multi-channel reverse dictionary model.

Michael Zock. 2004. Word lookup as an ongoing dialogue between a user and a lexicon. In *Proceedings of the 10th Annual Meeting of the Association for Natural Language Processing*, pages 484–487.

Michael Zock, Olivier Ferret, and Didier Schwab. 2010. Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, 13(4):201–218.