

Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis

Mario Giulianelli[◀], Iris Luden[◀], Raquel Fernández[◀], Andrey Kutuzov[◊]

[◊]University of Amsterdam [◊]University of Oslo

m.giulianelli@uva.nl, irisluden@gmail.com,
raquel.fernandez@uva.nl, andreku@ifi.uio.no

Abstract

We propose using automatically generated natural language definitions of contextualised word usages as interpretable word and word sense representations. Given a collection of usage examples for a target word, and the corresponding data-driven usage clusters (i.e., word senses), a definition is generated for each usage with a specialised Flan-T5 language model, and the most prototypical definition in a usage cluster is chosen as the sense label. We demonstrate how the resulting sense labels can make existing approaches to semantic change analysis more interpretable, and how they can allow users—historical linguists, lexicographers, or social scientists—to explore and intuitively explain diachronic trajectories of word meaning. Semantic change analysis is only one of many possible applications of the ‘definitions as representations’ paradigm. Beyond being human-readable, contextualised definitions also outperform token or usage sentence embeddings in word-in-context semantic similarity judgements, making them a new promising type of lexical representation for NLP.

1 Introduction

Accurate semantic understanding in language technologies is typically powered by distributional word representations and pre-trained language models (LMs). Due to their subsymbolic nature, however, such methods lack in explainability and interpretability, leading to insufficient trust in end users. An example application which requires capturing word meaning with its nuanced context-determined modulations is *lexical semantic change analysis*, a task which consists in detecting whether a word’s meaning has changed over time, for example by acquiring or losing a sense. Modern semantic change detection systems rely on static and contextualised word representations, LM-based lexical replacement, grammatical profiles, supervised word sense and word-in-context disambiguation (Kutuzov et al., 2018; Tahmasebi et al.,

2021). But the main potential end users of these technologies—historical linguists, lexicographers, and social scientists—are still somewhat reluctant to adopt them precisely because of their lack of explanatory power. Lexicographers, for instance, are not satisfied with detecting that a word has or hasn’t changed its meaning over the last ten years; they want descriptions of old and new senses in human-readable form, possibly accompanied by additional layers of explanation, e.g., specifying the type of semantic change (such as broadening, narrowing, and metaphorisation) the word has undergone.

Our work is an attempt to bridge the gap between computational tools for semantic understanding and their users. We propose to replace black-box contextualised token embeddings produced by large LMs with a new type of interpretable lexical semantic representation: automatically generated *contextualised word definitions* (Gardner et al., 2022). In this paradigm, the usage of the word ‘apple’ in the sentence ‘She tasted a fresh green apple’ is represented not with a dense high-dimensional vector but with the context-dependent natural language definition ‘EDIBLE FRUIT’. With an extended case study on lexical semantic change analysis, we show that moving to the more abstract meaning space of definitions allows practitioners to obtain explainable predictions from computational systems, while leading to superior performance on semantic change benchmarks compared to state-of-the-art token-based approaches.

This paper makes the following contributions.¹

1. We show that word definitions automatically generated with a specialised language model, fine-tuned for this purpose, can serve as interpretable representations for polysemous words (§5). Pairwise usage similarities between contextualised definitions approximate human semantic similarity judgements better

¹All the code we used can be found at https://github.com/lgtoslo/definition_modeling.

Usage example	Target word	Generated definition
‘about half of the soldiers in our rifle platoons were draftees whom we had trained for about six weeks’	draftee	‘A PERSON WHO IS BEING ENLISTED IN THE ARMED FORCES’

Table 1: An example of a definition generated by our fine-tuned Flan-T5 XL. The model is prompted with the usage example, post-fixed with the phrase ‘What is the definition of draftee?’

than similarities between usage-based word and sentence embeddings.

2. We present a method to obtain *word sense representations* by labelling data-driven clusters of word usages with sense definitions, and collect human judgements of definition quality to evaluate these representations (§6). We find that sense labels produced by retrieving the most prototypical contextualised word definition within a group of usages consistently outperform labels produced by selecting the most prototypical token embedding.
3. Using sense labels obtained via definition generation, we create maps that describe diachronic relations between the senses of a target word. We then demonstrate how these *diachronic maps* can be used to explain meaning changes observed in text corpora and to find inconsistencies in data-driven groupings of word usages within existing lexical semantic resources (§7).

2 Related Work

2.1 Definition Modelling

The task of generating human-readable word definitions, as found in dictionaries, is commonly referred to as *definition modelling* or *definition generation* (for a review, see Gardner et al., 2022). The original motivation for this task has been the interpretation, analysis, and evaluation of word embedding spaces. Definition generation systems, however, also have practical applications in lexicography, language acquisition, sociolinguistics, and within NLP (Bevilacqua et al., 2020). The task was initially formulated as the generation of a natural language definition given an embedding—a single distributional representation—of the target word, or *definiendum* (Noraset et al., 2017). Word meaning, however, varies according to the context in which a word is used. This is particularly true for polysemous words, which can be defined in multiple, potentially very different ways depending on their context. The first formulation of definition

modelling was therefore soon replaced by the task of generating a contextually appropriate word definition given a target word embedding and an example usage (Gadetsky et al., 2018; Mickus et al., 2022). When the end goal is not the evaluation of embedding spaces, generating definitions from vector representations is still not the most natural formulation of definition modelling. Ni and Wang (2017) and Mickus et al. (2019) treat the task as a sequence-to-sequence problem: given an input sequence with a highlighted word, generate a contextually appropriate definition. In the current work, we follow this approach. Table 1 shows an example of a contextualised definition generated by our model (see §4) for the English word ‘draftee’.

Methods Methods that address this last formulation of the task are typically based on a pre-trained language model deployed on the definienda of interest in a natural language generation (NLG) setup (Bevilacqua et al., 2020). Generated definitions can be further improved by regulating their degree of specificity via specialised LM modules (Huang et al., 2021), by adjusting their level of complexity using contrastive learning training objectives (August et al., 2022), or by supplementing them with definitional sentences extracted directly from a domain-specific corpus (Huang et al., 2022). We will compare our results to the specificity-tuned T5-based text generator proposed by Huang et al. (2021).

Evaluation Generated definitions are typically evaluated with standard NLG metrics such as BLEU, NIST, ROUGE-L, METEOR or Mover-Score (e.g., Huang et al., 2021; Mickus et al., 2022), using precision@k on a definition retrieval task (Bevilacqua et al., 2020), or measuring semantic similarity between sentence embeddings obtained for the reference and the generated definition (Kong et al., 2022). Because reference-based methods are inherently flawed (for a discussion, see Mickus et al., 2022), qualitative evaluation is almost always presented in combination with these quantitative metrics. In this paper, we evaluate

generated definitions with automatic metrics and by collecting human judgements.

2.2 Semantic Change Detection

Words in natural language change their meaning over time; these diachronic processes are of interest both for linguists and NLP practitioners. Lexical semantic change detection (LSCD) is nowadays a well represented NLP task, with workshops (Tahmasebi et al., 2022) and several shared tasks (e.g., Schlechtweg et al., 2020; Kurtyigit et al., 2021). LSCD is usually cast either as binary classification (whether the target word changed its meaning or not) or as a ranking task (ordering target words according to the degree of their change). To evaluate existing approaches, manually annotated datasets are used: so-called DWUGs are described below in §3.

An important issue with current LSCD methods is that they rarely describe change in terms of *word senses*, which are extremely important for linguists to understand diachronic meaning trajectories. Instead, systems provide (and are evaluated by) per-word numerical ‘change scores’ which are hardly interpretable; at best, a binary ‘sense gain’ or ‘sense loss’ classification is used. Even approaches that do operate on the level of senses (e.g., Mitra et al., 2015; Homskiy and Arefyev, 2022) do not label them in a linguistically meaningful way, making it difficult to understand the relations between the resulting ‘anonymous’ types of word usage.

3 Data

3.1 Definitions Datasets

To train an NLG system that produces definitions (§4), we use three datasets containing a human-written definition for each lexicographic sense of a target word, paired with a usage example. The **WordNet** dataset is a collection of word definitions and word usages extracted by Ishiwatari et al. (2019) from the WordNet lexical database (Miller, 1995). The **Oxford** dataset (also known as CHA in prior work) consists of definitions and usage ex-

Dataset	Entries	Lemmas	Ratio	Usage length	Definition length
WordNet	15,657	8,938	1.75	4.80 ± 3.43	6.64 ± 3.77
Oxford	122,318	36,767	3.33	16.73 ± 9.53	11.01 ± 6.96
CoDWoE	63,596	36,068	2.44	24.04 ± 21.05	11.78 ± 8.03

Table 2: Main statistics of the datasets of definitions. Ratio is the *sense-lemma* ratio: the number of entries over the number of lemmas.

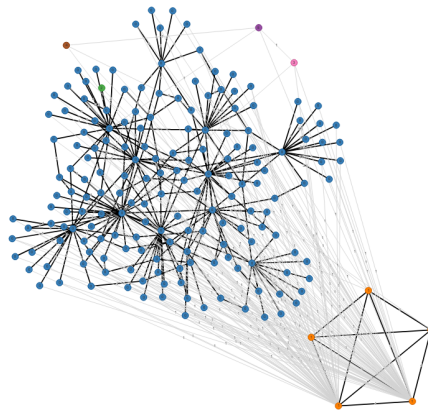


Figure 1: Diachronic word usage graph for the English word ‘lass’ (Schlechtweg et al., 2021).

amples collected by Gadetsky et al. (2018) from the Oxford Dictionary. Definitions are written by experts and usage examples are in British English. The **CoDWoE** dataset (Mickus et al., 2022) is based on definitions and examples extracted from Wiktionary.² It is a multilingual corpus, of which we use the English portion. Table 2 reports the main statistics of these datasets. Further statistics, e.g. on the size of the different splits, are provided by Huang et al. (2021) as well as in Appendix A.³

3.2 Diachronic Word Usage Graphs

We showcase interpretable word usage (§5) and sense representations (§6 and 7) using a dataset where target lemmas are represented with diachronic word usage graphs (DWUGs, Schlechtweg et al., 2021). A DWUG is a weighted, undirected graph, where nodes represent target usages (word occurrences within a sentence or discourse context) and edge weights represent the semantic proximity of a pair of usages. DWUGs are the result of a multi-round incremental human annotation process, with annotators asked to judge the semantic relatedness of pairs of word usages on a 4-point scale. Based on these pairwise relatedness judgements, word usages are then grouped into usage clusters (a data-driven approximation of *word senses*) using a variation of correlation clustering (Bansal et al., 2004; Schlechtweg et al., 2020).

DWUGs are currently available in seven

²<https://www.wiktionary.org>

³Note that in theory, a definition dataset could be also be extracted from the SemCor corpus (Miller et al., 1993). However, we do not anticipate it will contribute much to training or evaluation since SemCor does not contain any new definitions with respect to WordNet: only more examples for the same word-definition pairs. This can be investigated in future work.

languages.⁴ In this paper, we use the English graphs, which consist of usage sentences sampled from the Clean Corpus of Historical American English (Davies, 2012; Alatrash et al., 2020) and belonging to two time periods: 1810-1860 and 1960-2010. There are 46 usage graphs for English, corresponding to 40 nouns and 6 verbs annotated by a total of 9 annotators. Each target lemma has received on average 189 judgements, 2 for each usage pair. Figure 1 shows an example of a DWUG, with colours denoting usage clusters (i.e., data-driven senses): the ‘blue’ and ‘orange’ clusters belong almost entirely to different time periods: a new sense of the word has emerged. We show how our approach helps explain such cases of semantic change in §7.

4 Definition Generation

Our formulation of the *definition generation* task is as follows: given a target word w and an example usage s (i.e., a sentence containing an occurrence of w), generate a natural language definition d that is grammatical, fluent, and faithful to the meaning of the target word w as used in the example usage s . A *definition generator* is a language process that maps words and example usages to natural language definitions. As a generator, we use Flan-T5 (Chung et al., 2022), a version of the T5 encoder-decoder Transformer (Raffel et al., 2020) fine-tuned on 1.8K tasks phrased as instructions and collected from almost 500 NLP datasets. Flan-T5 is not trained specifically on definition generation but thanks to its massive multi-task instruction fine-tuning, the model exhibits strong generalisation to unseen tasks. Therefore, we expect it to produce high-quality definitions. We extensively test three variants of Flan-T5 of different size and compare them to vanilla T5 models (Table 4 and Table 12, Appendix C.2); based on our results, we recommend using the largest fine-tuned Flan-T5 model whenever possible.

To obtain definitions from Flan-T5, we use natural language prompts consisting of an example usage preceded or followed by a question or instruction. For example: ‘ s What is the definition of w ?’ The concatenated usage example and prompt are provided as input to Flan-T5, which conditionally generates definitions (Ta-

ble 1 shows an example instance).⁵ We choose greedy search with target word filtering as a simple, parameter-free decoding strategy. Stochastic decoding algorithms can be investigated in future work.

Prompt selection In preliminary experiments, we used the pre-trained Flan-T5 Base model (250M parameters) to select a definition generation prompt among 8 alternative verbalisations. Appending the question ‘*What is the definition of w ?*’ to the usage example consistently yielded the best scores.⁶ We use this prompt for all further experiments.

4.1 Evaluating Generated Definitions

Before using its definitions to construct an interpretable semantic space—the main goal of this paper—we perform a series of experiments to validate Flan-T5 as a definition generator. We use the target lemmas and usage examples from the corpora of definitions presented in §3, conditionally generate definitions with Flan-T5, and then compare them to the gold definitions in the corpora using reference-based NLG evaluation metrics. We report SacreBLEU and ROUGE-L, which measure surface form overlap, as well as BERT-F1, which is sensitive to the reference and candidate’s semantics. As mentioned in §2.1, reference-based metrics are not flawless, yet designing and validating a reference-free metric for the definition generation task is beyond the scope of this paper. We will later resort to correlations with human judgements and expert human evaluation to assess the quality of generated definitions.

We evaluate the Flan-T5 XL (3B parameters) in three generalisation tests: 1) in distribution, 2) hard domain shift, and 3) soft domain shift.⁷ We use these tests to choose a model to be deployed in further experiments. For reference, we report the BLEU score of the definition generator by Huang et al. (2021); ROUGE-L and BERT-F1 are not reported in their paper.

In distribution We fine-tune Flan-T5 XL on one corpus of definitions at a time, and test it on a held-out set from that same corpus (except

⁴<https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/>

⁵This is a simpler workflow in comparison to prior work (Bevilacqua et al., 2020; Almeman and Espinosa Anke, 2022) where inputs are encoded as ‘target word - context’ pairs.

⁶Further details in Appendix B.

⁷Tests defined following the GenBench generalisation taxonomy (Hupkes et al., 2022). We also include a fourth setup, *zero shot (task shift)*, where we directly evaluate the pretrained Flan-T5 XL. Results (including other models) are presented in Appendix C.1-C.2, and an evaluation card in Appendix C.3.

Model	Test	WordNet			Oxford		
		BLEU	ROUGE-L	BERT-F1	BLEU	ROUGE-L	BERT-F1
Huang et al. (2021)	Unknown	32.72	-	-	26.52	-	-
Flan-T5 XL	Zero-shot (task shift)	2.70	12.72	86.72	2.88	16.20	86.52
Flan-T5 XL	In-distribution	11.49	28.96	88.90	16.61	36.27	89.40
Flan-T5 XL	Hard domain shift	29.55	48.17	91.39	8.37	25.06	87.56
Flan-T5 XL	Soft domain shift	32.81	52.21	92.16	18.69	38.72	89.75

Table 3: Results of the definition generation experiments.

CoDWoE which does not provide train-test split). The quality of the definitions increases substantially with fine-tuning, in terms of both their lexical and semantic overlap with gold definitions (Table 3). We find significantly higher scores on Oxford, which may be due to the larger size of its training split and to the quality of the WordNet examples, which sometimes are not sufficiently informative (Almeman and Espinosa Anke, 2022).

Hard domain shift We fine-tune Flan-T5 XL on WordNet and test it on Oxford, and vice versa. These tests allow us to assess the model’s sensitivity to the peculiarities of the training dataset. A model that has properly learned to generate definitions should be robust to this kind of domain shift. The quality of the definitions of Oxford lemmas generated with the model fine-tuned on WordNet (see the Oxford column in Table 3) is lower than for the model fine-tuned on Oxford itself (same column, see row ‘In-distribution’). Instead, for out-of-domain WordNet definitions, all metrics surprisingly indicate higher quality than for in-distribution tests (WordNet column). Taken together, our results so far suggest that the quality of a fine-tuned model depends more on the amount (and perhaps quality) of the training data than on whether the test data is drawn from the same dataset.

Soft domain shift We finally fine-tune Flan-T5 XL on a collection of all three definition datasets: WordNet, Oxford, and CoDWoE. Our previous results hint towards the model’s preference for more training examples, so we expect this setup to achieve the highest scores regardless of the soft shift between training and test data. Indeed, on WordNet, our fine-tuned model marginally surpasses the state-of-the-art upper bound in terms of BLEU score (Table 3), and it achieves the highest scores on the other metrics. Oxford definitions generated with this model are instead still below Huang et al.’s upper bound; this may be due to Oxford being generally more difficult to model

than WordNet, perhaps because of longer definitions and usages (see Figures 4-5 in Appendix A). We consider the observed model performance sufficient for the purposes of our experiments, in particular in view of the higher efficiency of fine-tuned Flan-T5 with respect to the three-module system of Huang et al. (2021). We therefore use this model throughout the rest of our study.

The Flan-T5 models fine-tuned for definition generation are publicly available through the Hugging Face model hub.⁸

5 Definitions are Interpretable Word Representations

We propose considering the abstract meaning space of definitions as a representational space for lexical meaning. Definitions fulfil important general desiderata for word representations: they are human-interpretable and they can be used for quantitative comparisons between word usages (i.e., by judging the distance between pairs of definition strings). We put the *definition space* to test by applying it to the task of semantic change analysis, which requires capturing word meaning at a fine-grained level, distinguishing word senses based on usage contexts. We use our fine-tuned Flan-T5 models (XL and other sizes) to generate definitions for all usages of the 46 target words annotated in the English DWUGs (ca. 200 usages per word; see §3.2).⁹ These definitions (an example is provided in Table 1) specify a diachronic semantic space.

5.1 Correlation with Human Judgements

We construct word usage graphs for each lemma in the English DWUGs: we take usages as nodes and assign weights to edges by measuring pairwise similarity between usage-dependent definitions. We

⁸Model names: [lgt/flan-t5-definition-en-base](#), [lgt/flan-t5-definition-en-large](#), [lgt/flan-t5-definition-en-xl](#).

⁹The training datasets used in §4 contain nouns, verbs, adjectives and adverbs. The English DWUGs contain only nouns and verbs.

Method	Cosine	SacreBLEU	METEOR
Token embeddings	0.141	-	-
Sentence embeddings	0.114	-	-
Generated definitions			
FLAN-T5 XL Zero-shot	0.188	0.041	0.083
FLAN-T5 XXL Zero-shot	0.206	0.045	0.092
FLAN-T5 base FT	0.221	0.078	0.077
FLAN-T5 XL FT	0.264	0.108	0.117

Table 4: Correlations with pairwise similarity judgements by humans. ‘FT’ stands for ‘fine-tuned model’.

compute the similarity between pairs of definitions using two overlap-based metrics, SacreBLEU and METEOR, as well as the cosine similarity between sentence-embedded definitions. We then compare our graphs against the gold DWUGs, where edges between usage pairs are weighted with human judgements of semantic similarity, by computing the Spearman’s correlation between human similarity judgements and similarity scores obtained for pairs of generated definitions. We compare our results to DWUGs constructed based on two additional types of usage-based representations: *sentence* embeddings obtained directly for usage examples, and contextualised *token* embeddings. Sentence embeddings (for both definitions and usage examples) are SBERT representations (Reimers and Gurevych, 2019) extracted with mean-pooling from the last layer of a DistilRoBERTa LM fine-tuned for semantic similarity comparisons.¹⁰ For tokens, we extract the last-layer representations of a RoBERTa-large model (Liu et al., 2019) which correspond to subtokens of the target word (following Giulianelli et al., 2020) and use mean-pooling to obtain a single vector. While we report string-overlap similarities for definitions, these are not defined for numerical vectors, and thus similarities for example sentences and tokens are obtained with cosine only.

Pairwise similarities between definitions approximate human similarity judgements far better than similarities between example sentence and word embeddings (Table 4). This indicates that definitions are a more accurate approximation of contextualised lexical meaning. The results also show that similarity between definitions is best captured by their embeddings, rather than by overlap-based

¹⁰DistilRoBERTa (sentence-transformers/all-distilRoBERTa-v1) is the second best model as reported in the official S-BERT documentation at the time of publication (https://www.sbert.net/docs/pretrained_models.html). For a negligible accuracy reduction, it captures longer context sizes and is ca. 50% smaller and faster than the model that ranks first.

metrics like SacreBLEU and METEOR.

5.2 Definition Embedding Space

We now examine the *definition embedding space* (the high-dimensional semantic space defined by sentence-embedded definitions), to identify properties that make it more expressive than usage-based spaces. Figure 2 shows the T-SNE projections of the DistilRoBERTa embeddings of all lemmas in the English DWUGs, for the three types of representation presented earlier: generated definitions, tokens, and example sentences.¹¹ The definition spaces exhibit characteristics that are more similar to a *token* embedding space than an example *sentence* embedding space, with definitions of the same lemma represented by relatively close-knit clusters of definition embeddings. This suggests that definition embeddings, as expected, represent the meaning of a word in context (similar to token embeddings), rather than the meaning of the whole usage example sentence in which the target word occurs.

For each target word, we also measure (i) the variability in each embedding space and (ii) the inter-cluster and intra-cluster dispersion (Caliński and Harabasz, 1974) obtained when clustering each space using k -means. This allows us to quantitatively appreciate properties exhibited by data-driven usage clusters that are obtained from different representation types. To cluster the embedding spaces, we experiment with values of $k \in [2, 25]$, and select the k which maximises the Silhouette score. Our results are summarised in Table 5. We observe that the clusters in the definition spaces have on average the lowest intra-cluster dispersion, indicating that they are more cohesive than the clusters in the token and example sentence spaces. While, on average, token spaces exhibit higher inter-cluster dispersion (indicating better cluster separation), the ratio between average separation and cohesion is highest for the definition spaces. These findings persist for the gold clusters determined by the English DWUGs (Table 14).

In sum, this analysis shows that definition embedding spaces are generally suitable to distinguish different types of word usage. In the next section, we will show how they can indeed be used to characterise word senses.

¹¹T-SNE projections for RoBERTa-large are in Appendix G.



Figure 2: T-SNE projection of each embedding space, DistilRoBERTa model.

Model	Representation	Variance	Std	K	Silh.	Sep.	Coh.	Ratio
RoBERTa-large	Sentence	0.014	0.117	2.0	0.111	0.285	0.012	23.2
	Token	0.034	0.183	3.8	0.173	0.868	0.027	32.4
	Definitions	0.006	0.080	20.6	0.335	0.057	0.003	22.3
DistilRoBERTa	Sentence	0.597	0.772	2.1	0.046	4.907	0.578	8.5
	Token	0.477	0.687	2.5	0.121	8.599	0.427	20.1
	Definitions	0.509	0.756	19.7	0.355	5.559	0.228	24.4

Table 5: Variance, standard deviation, optimal K , silhouette score, separation score, cohesion score, and the separation-cohesion ratio for each embedding space; average over all target words.

6 Labelling Word Senses With Definitions

For generated definitions to be useful in practice, they need to be able to distinguish word senses. For example (ignoring diachronic differences and singleton clusters), there are three main senses of the word ‘word’ in its DWUG, which we manually label as: (1) ‘WORDS OF LANGUAGE’, (2) ‘A RUMOUR’, and (3) ‘AN OATH’. Manual inspection of the generated definitions indicates that they are indeed sense-aware:

1. ‘A communication, a message’, ‘The text of a book, play, movie’, etc.
2. ‘Information passed on, usually by one person to another’, ‘communication by spoken or written communication’, etc.
3. ‘An oath’, ‘a pronouncement’, etc.

But let’s again put ourselves in the shoes of a historical linguist. Sense clusters are now impractically represented with multitudes of contextualised definitions. Cluster (1) for ‘word’, e.g., features 190 usages, and one must read through all of them (otherwise there will be a chance of missing something) and generalise – all to formulate a definition that covers the whole sense cluster (a *sense label*). We now show how DWUGs can be automatically augmented with generated sense labels, vastly improving their usability.

Selecting sense labels From n definitions, generated for n word usages belonging to the same DWUG cluster, we use the most prototypical one

as the *sense label*—with the aim of reflecting the meaning of the majority of usages in the cluster. We represent all definitions with their sentence embeddings (cf. §5.1) and select as prototypical the definition whose embedding is most similar to the average of all embeddings in the cluster. Clusters with less than 3 usages are ignored as, for these, prototypicality is ill-defined. As a sanity check, these are the sense labels obtained by this method for the DWUG clusters of ‘word’; they correspond well to the sense descriptions provided earlier.

1. ‘A SINGLE SPOKEN OR WRITTEN UTTERANCE’
2. ‘INFORMATION; NEWS; REPORTS’
3. ‘A PROMISE, VOW OR STATEMENT’

We compare these sense labels to labels obtained by generating a definition for the most prototypical *usage* (as judged by its token embedding), rather than taking the most prototypical *definition*, and we evaluate both types of senses labels using human judgements. Examples of labels can be found in Appendix D.

Human evaluation Five human annotators (fluent English speakers) were asked to evaluate the quality of sense labels for each cluster in the English DWUGs, 136 in total. Each cluster was accompanied by the target word, two labels (from definitions and from usages) and five example usages randomly sampled from the DWUG. The annotators could select one of six judgements to indicate overall quality of the labels and their relative ranking. After a reconciliation round, the categorical judgements were aggregated via majority voting. Krippendorff’s α inter-rater agreement is 0.35 on the original data and 0.45 when the categories are reduced to four. Full guidelines and results are reported in Appendix E.¹²

We find that our prototypicality-based sense labelling strategy is overall reliable. Only for 15% of the clusters, annotators indicate that neither

¹²There exist no established procedures for the collection of human quality judgements of automatically generated word sense labels. The closest efforts we are aware of are those in Noraset et al. (2017), who ask annotators to rank definitions generated by two systems, providing as reference the gold dictionary definitions. In our case, (1) generations are for word senses rather than lemmas, (2) we are interested not only in rankings but also in judgements of ‘sufficient quality’, (3) dictionary definitions are not available for the DWUG senses; instead (4) we provide annotators with usage examples, which are crucial for informed judgements of sense definitions.

of the labels is satisfactory (Figure 9). When comparing definition-based and usage-based labels, the former were found to be better in 31% of the cases, while the latter in only 7% (in the rest of the cases, the two methods are judged as equal). We also analysed how often the labels produced by each method were found to be acceptable. Definition-based labels were of sufficient quality in 80% of the instances, while for usage-based labels this is only true for 68% of the cases.

In sum, prototypical definitions reflect sense meanings better than definitions of prototypical usage examples. We believe this is because definitions are more abstract and robust to contextual noise (the same definition can be assigned to very different usages, if the underlying sense is similar). This approach takes the best of both worlds: the produced representations are data-driven, but at the same time they are human-readable and naturally explanatory. After all, ‘senses are abstractions from clusters of corpus citations’ (Kilgarriff, 1997). In the next section, we demonstrate how automatically generated definition-based sense labels can be used to explain semantic change observed in diachronic text corpora.

7 Explaining Semantic Change with Sense Labels

Word senses in DWUGs are collections of example usages and they are only labelled with numerical identifiers. This does not allow users to easily grasp the meaning trajectories of the words they are interested in studying. Using sense labels extracted from generated definitions, we can produce a fully human-readable *sense dynamics map*—i.e., an automatically annotated version of a DWUG which displays synchronic and diachronic relations between senses (e.g. senses transitioning one into another, splitting from another sense, or two senses merging into one). One can look at sense dynamics maps as reproducing the work of Mitra et al. (2015) on the modern technological level and, importantly, with human-readable sense definitions.

Given a target word, its original DWUG, and its semi-automatic sense clusters, we start by assigning a definition label to each cluster, as described in §6. Then, we divide each cluster into two sub-clusters, corresponding to time periods 1 and 2 (for example, sub-cluster c_1^2 contains all usages from cluster 1 occurring in time period 2).¹³ We

¹³Note that the labels are still time-agnostic: that is, sub-

compute pairwise cosine similarities between the sentence embeddings of the labels (their ‘definition embeddings’), thereby producing a fully connected graph where nodes are sub-clusters and edges are weighted with sense label similarities. Most edges have very low weight, but some sub-cluster pairs are unusually similar, hinting at a possible relation between the corresponding senses. We detect these outlier pairs by inspecting the distribution of pairwise similarities for values with z -score higher than 1 (similarities more than 1 standard deviation away from the mean similarity). Sub-cluster pairs connected with such edges form a *sense dynamics map*.

As an example, the noun ‘*record*’ has only one sense in time period 1 but it acquires two new senses in time period 2 (Figure 3; as before, we ignore clusters with less than 3 usages). The sense clusters defined by the DWUG are anonymous collection of usages, but with the assigned sense labels (also shown in Figure 3) they can be turned into a proto-explanation of the observed semantic shift:

- A novel sense 2 of ‘*record*’ in time period 2 (‘A PHONOGRAPH OR GRAMOPHONE CYLINDER CONTAINING AN AUDIO RECORDING.’) is probably an offshoot of a stable sense 0 present in both time periods (‘A DOCUMENT OR OTHER MEANS OF PROVIDING INFORMATION ABOUT PAST EVENTS.’).

It becomes now clear that sense 2 stems from the older general sense 0 of ‘*record*’—arguably representing a case of narrowing (Bloomfield, 1933)—while the second new sense (1: ‘THE HIGHEST SCORE OR OTHER ACHIEVEMENT IN THE GAME’) is not related to the others and is thus independent.

Sense dynamics maps can also help in tracing potentially incorrect or inconsistent clustering in DWUGs. For instance, if different sense clusters are assigned identical definition labels, then it is likely that both clusters correspond to the same sense and that the clustering is thus erroneous. Using our automatically produced sense dynamics maps, DWUGs can be improved and enriched (semi-)automatically.

An interesting case is ‘*ball*’ (see Appendix F for another example regarding the word ‘*chef*’).

clusters c_1^1 and c_1^2 have the same label. This is done for simplicity and because of data scarcity, but in the future we plan to experiment with time-dependent labels as well. We use two time periods as only two periods are available in Schlechtweg et al.’s English DWUGs (2021), but the same procedure can be executed on multi-period datasets.

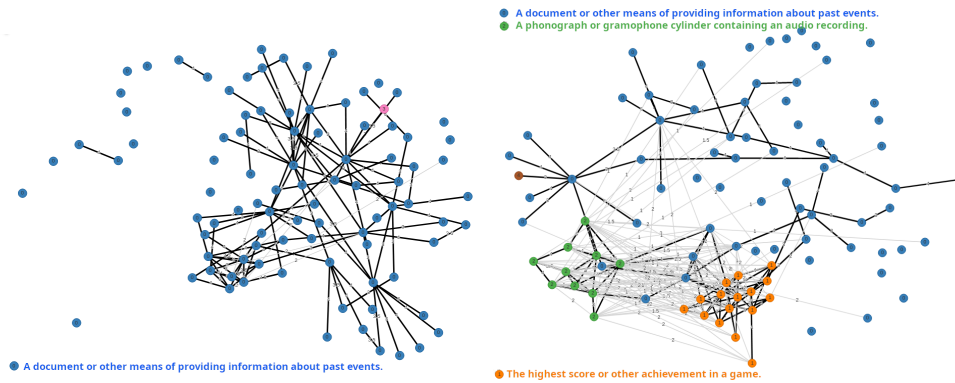


Figure 3: Diachronic word usage graphs for ‘record’ (Schlechtweg et al., 2021) with sense definitions generated using our proposed procedure (§6). Left: time period 1 (1810-1860); right: time period 2 (1960-2010). Colours correspond to data-driven senses, as annotated in the original DWUGs.

Although none of its sense labels are identical, its sense cluster c_0 is very close to cluster c_2 (similarity of 0.70), while c_2 is close to c_3 (similarity of 0.53); all three senses persist throughout both time periods, with sense 3 declining in frequency. The generated definitions for the ‘ball’ clusters are: 0: ‘A SPHERE OR OTHER OBJECT USED AS THE OBJECT OF A HIT’ (the largest cluster), 2: ‘A ROUND SOLID PROJECTILE, SUCH AS IS USED IN SHOOTING’, and 3: ‘A BULLET’. This case demonstrates that similarity relations are not transitive: the similarity between c_0 and c_3 is only 0.50, below our outlier threshold value. This is in part caused by inconsistent DWUG clustering: while the majority of usages in c_2^1 are about firearm projectiles, c_2^2 contains mentions of golf balls and ball point pens. This shifts sense 2 from ‘BULLET’ to ‘ROUND SOLID PROJECTILE’, making it closer to sense 0 (general spheres) than it should be. Ideally, all the ‘BULLET’ usages from c_2 should have ended up in c_3 , with the rest joining the general sense 0.

Besides suggesting fixes to the DWUG clustering, the observed non-transitivity also describes a potential (not necessarily diachronic) meaning trajectory of ‘ball’: from any spherical object, to spherical objects used as projectiles, and then to any projectiles (like bullets), independent of their form. Our generated sense labels and their similarities help users analyse this phenomenon in a considerably faster and easier way than by manually inspecting all examples for these senses.

8 Conclusion and Future Work

In this paper, we propose to consider automatically generated contextualised word definitions as a type of lexical representation, similar to traditional

word embeddings. While generated definitions have been already shown to be effective for word sense disambiguation (Bevilacqua et al., 2020), our study puts this into a broader perspective and demonstrates that modern language models like Flan-T5 (Chung et al., 2022) are sufficiently mature to produce robust and accurate definitions in a simple prompting setup. The generated definitions outperform traditional token embeddings in word-in-context similarity judgements while being naturally interpretable.

We apply definition-based lexical representations to semantic change analysis and show that our approach can be used to trace word sense dynamics over time. Operating in the space of human-readable definitions makes such analyses much more interesting and actionable for linguists and lexicographers—who look for explanations, not numbers. At the same time, we believe the ‘definitions as representations’ paradigm can also be used for other NLP tasks in the area of lexical semantics, such as word sense induction, idiom detection, and metaphor interpretation.

Our experiments with diachronic sense modelling are still preliminary and mostly qualitative. It is important to evaluate systematically how well our predictions correspond to the judgements of (expert) humans. Once further evidence is gathered, other promising applications include tracing cases of semantic narrowing or widening over time (Bloomfield, 1933) by analysing the variability of contextualised definitions in different time periods and by making cluster labels time-dependent. Both directions will require extensive human annotation, and we leave them for future work.

Limitations

Data in this work is limited to the English diachronic word usage graphs (DWUGs). Our methods themselves are language-agnostic and we do not anticipate serious problems with adapting them to DWUGs in other languages (which already exist). At the same time, although Flan-T5 is a multilingual LM, we did not thoroughly evaluate its ability to generate definitions in languages other than English. Again, definition datasets in other languages do exist and technically it is trivial to fine-tune Flan-T5 on some or all of them.

Generated definitions and mappings between definitions and word senses can contain all sorts of biases and stereotypes, stemming from the underlying language model. Filtering inappropriate character strings from the definitions can only help as much, and further research is needed to estimate possible threats.

In our experiments with Flan-T5, the aim was to investigate the principal possibility of using this LM for definition modelling. Although we did evaluate several different Flan-T5 variants, we leave it for the future work to investigate the impact of model size and other experimental variables (such as decoding algorithms).

The cases shown in §7 are hand-picked examples, demonstrating the potential of using generated definitions for explainable semantic change detection and improving LSCD datasets. In the future, we plan to conduct a more rigorous evaluation of different ways to build sense dynamics map.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819455). The computations were performed on resources provided through Sigma2—the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. **CCOHA: Clean corpus of historical American English**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

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, Taipei, Taiwan. Association for Computational Linguistics.

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.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. **Correlation clustering**. *Machine Learning*, 56(1):89–113.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. **Generatory 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.

Leonard Bloomfield. 1933. *Language*. Allen & Unwin.

Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. **Scaling instruction-finetuned language models**. *arXiv preprint arXiv:2210.11416*.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

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.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. **Definition modeling: Literature review and dataset analysis**. *Applied Computing and Intelligence*, 2(1):83–98.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. **Analysing lexical semantic change with contextualised word representations**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

- Daniil Homskiy and Nikolay Arefyev. 2022. [DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators?](#) In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. [Understanding jargon: Combining extraction and generation for definition modeling.](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2022. State-of-the-art generalisation research in NLP: A taxonomy and review. *arXiv preprint arXiv:2210.03050*.
- 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.
- Adam Kilgarriff. 1997. I don’t believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. [Multitasking framework for unsupervised simple definition generation.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.
- Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. [Lexical semantic change discovery.](#) 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 6985–6998, Online. Association for Computational Linguistics.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey.](#) In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- 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*.
- Timothee Mickus, Denis Paperno, and Matthieu Constanant. 2019. [Mark my word: A sequence-to-sequence approach to definition modeling.](#) In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.
- Timothee Mickus, Kees Van Deemter, Mathieu Constanant, 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.
- George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. 1993. [A semantic concordance.](#) In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.
- 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. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. *SemEval-2020 task 1: Unsupervised lexical semantic change detection*. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. *DWUG: A large resource of diachronic word usage graphs in four languages*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6:1.

Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors. 2022. *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Dublin, Ireland.

Appendix

A Preliminary Analysis of Usage Examples

In Section 3.1 of the main paper, we present three corpora of human-written definitions and report their main statistics in Table 2, including mean and standard deviation of usage example length. Because the length of usage examples has been shown to affect the quality of generated definitions (Alman and Espinosa Anke, 2022), in a preliminary analysis, we compare the length distributions of usage examples in the corpora of definitions as well as in the English DWUGs (Schlechtweg et al., 2021). Figures 4–7 show the length distributions of the four datasets. We also measure the correlation between definition quality (BertScore, BLEU, NIST) and (i) the length of usage examples, (ii) the absolute position of the target word in the examples, and (iii) the target word’s relative position in the examples. Tables 6 and 7 show the correlation coefficients.

	Length	Relative Position	Absolute Position	BertScore	Bleu	Nist
Length	1.000000	-0.121793	0.575304	0.067180	0.076133	0.044873
Relative Position	-0.121793	1.000000	0.626032	0.052725	0.074697	0.062041
Absolute Position	0.575304	0.626032	1.000000	0.128785	0.159078	0.110559
BertScore	0.067180	0.052725	0.128785	1.000000	0.121067	0.095343
Bleu	0.076133	0.074697	0.159078	0.121067	1.000000	0.821956
Nist	0.044873	0.062041	0.110559	0.095343	0.821956	1.000000

Table 6: Correlations between properties of the usage examples and the quality (BertScore, BLEU, NIST) of the definitions generated by Flan-T5 Base for WordNet. The prompt used is ‘What is the definition of w ?’ (post). The maximum context size is set to 512.

	Length	Relative Position	Absolute Position	BertScore	Bleu	Nist
Length	1.000000	-0.040948	0.615536	0.019844	0.039525	0.017253
Relative Position	-0.040948	1.000000	0.674509	0.046071	0.019940	0.023542
Absolute Position	0.615536	0.674509	1.000000	0.029413	0.016901	0.006764
BertScore	0.019844	0.046071	0.029413	1.000000	0.283203	0.276626
Bleu	0.039525	0.019940	0.016901	0.283203	1.000000	0.687382
Nist	0.017253	0.023542	0.006764	0.276626	0.687382	1.000000

Table 7: Correlations between properties of the usage examples and the quality (BertScore, BLEU, NIST) of the definitions generated by Flan-T5 Base for Oxford. The prompt used is ‘What is the definition of w ?’ (post). The maximum context size is set to 512.

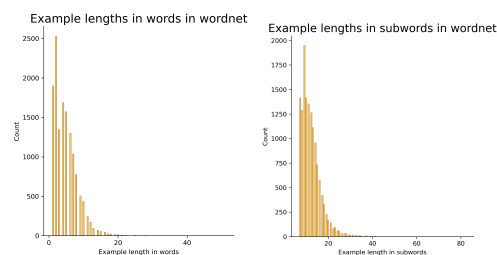


Figure 4: Length distribution of usage examples in WordNet.

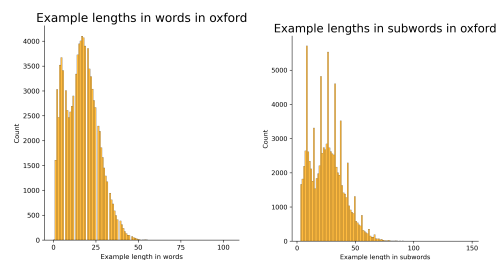


Figure 5: Length distribution of usage examples in Oxford.

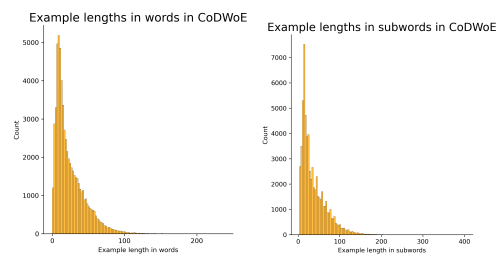


Figure 6: Length distribution of usage examples in CoD-WoE.

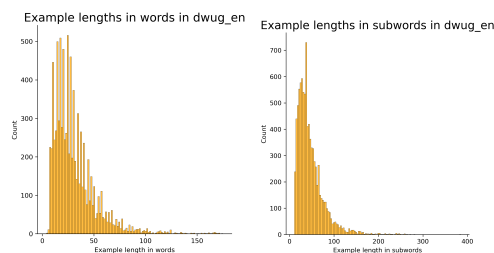


Figure 7: Length distribution of usage examples in the English DWUGs.

Configuration	BLEU	NIST	BERTScore
what is the definition of <trg>? post 256	0.0985	0.1281	0.8700
what is the definition of <trg>? post 512	0.0985	0.1281	0.8700
give the definition of <trg> post filter	0.0719	0.1520	0.8560
give the definition of <trg> post 256	0.0629	0.1563	0.8522
give the definition of <trg> post 512	0.0629	0.1563	0.8522
define the word <trg> post 512	0.0462	0.0972	0.8512
define the word <trg> post 256	0.0462	0.0972	0.8512
give the definition of <trg>: pre 256	0.0446	0.1123	0.8495
what is the definition of <trg>? pre 512	0.0403	0.0705	0.8495
give the definition of <trg>: pre 512	0.0446	0.1123	0.8495
what is the definition of <trg>? pre 256	0.0403	0.0703	0.8494
define the word <trg>: pre 512	0.0313	0.0615	0.8481
define the word <trg>: pre 256	0.0313	0.0618	0.8480
define <trg> post 512	0.0275	0.0583	0.8475
define <trg> post 256	0.0275	0.0583	0.8475
define <trg>: pre 512	0.0195	0.0411	0.8453
define <trg>: pre 256	0.0195	0.0409	0.8453

Table 8: Prompt selection results on WordNet (see description in Appendix B).

B Prompt Selection

As briefly discussed in Section 4, in preliminary experiments, we use the pretrained Flan-T5 Base model (250M parameters; Chung et al., 2022) to select a definition generation prompt among 8 alternative verbalisations. These are a combination of four different instruction strings (‘Define w ’, ‘Define the word w ’, ‘Give the definition of w ’, ‘What is the definition of w ’) and two ways of concatenating instructions to usage examples – i.e., either prepending them or appending them. Tables 8-11 show the results of our experiments. In the tables, the strings ‘pre’ and ‘post’ refer to the concatenation method (prepending or appending the instruction), the numbers 128, 256, and 512 refer to the maximum length of the usage examples provided to Flan-T5 (in sub-words), and ‘filter’ refers to the decoding strategy of always avoiding the target word (definiendum).

Configuration	BLEU	NIST	BERTScore
what is the definition of <trg>? post 512	0.1232	0.1488	0.8648
what is the definition of <trg>? post 128	0.1232	0.1488	0.8648
what is the definition of <trg>? post 256	0.1232	0.1488	0.8648
what is the definition of <trg>? post oxford filter 128	0.1219	0.1398	0.8644
give the definition of <trg> post 128	0.0823	0.1793	0.8531
give the definition of <trg> post 256	0.0823	0.1793	0.8531
give the definition of <trg> post 512	0.0823	0.1793	0.8531
give the definition of <trg> post oxford filter 128	0.0763	0.1415	0.8526
what is the definition of <trg>? pre 256	0.0801	0.0966	0.8501
what is the definition of <trg>? pre 512	0.0801	0.0966	0.8501
what is the definition of <trg>? pre 128	0.0801	0.0966	0.8501
give the definition of <trg>: pre 128	0.0695	0.1313	0.8493
give the definition of <trg>: pre 256	0.0695	0.1313	0.8493
give the definition of <trg>: pre 512	0.0695	0.1313	0.8492
define the word <trg> post 128	0.0614	0.1112	0.8442
define the word <trg> post 512	0.0614	0.1112	0.8442
define the word <trg> post 256	0.0614	0.1112	0.8442
define the word <trg>: pre 256	0.0408	0.0602	0.8352
define the word <trg>: pre 512	0.0408	0.0602	0.8352
define the word <trg>: pre 128	0.0408	0.0602	0.8352
define <trg> post 256	0.0279	0.0581	0.8319
define <trg> post 128	0.0279	0.0581	0.8319
define <trg> post 512	0.0279	0.0581	0.8319
define <trg>: pre 512	0.0161	0.0237	0.8305
define <trg>: pre 256	0.0160	0.0237	0.8305
define <trg>: pre 128	0.0160	0.0237	0.8305

Table 9: Prompt selection results on Oxford (see description in Appendix B).

Configuration	BLEU	NIST	BERTScore
what is the definition of <trg>? post 128	0.1138	0.2137	0.8702
give the definition of <trg> post 128	0.0826	0.2389	0.8615
what is the definition of <trg>? post 64	0.1033	0.1990	0.8595
give the definition of <trg> post 64	0.0785	0.2194	0.8520

Table 10: Prompt selection results on CoDWoE Complete (see description in Appendix B).

Configuration	BLEU	NIST	BERTScore
give the definition of <trg>: pre 64	0.0680	0.1513	0.8461
what is the definition of <trg>? post 64	0.1068	0.1464	0.8458
give the definition of <trg> post 64	0.0654	0.1602	0.8374

Table 11: Prompt selection results on CoDWoE Trial (see description in Appendix B).

Model	Test	WordNet			Oxford		
		BLEU	ROUGE-L	BERT-F1	BLEU	ROUGE-L	BERT-F1
Huang et al. (2021)	Unknown	32.72	-	-	26.52	-	-
T5 base	Zero-shot (task shift)	2.01	8.24	82.98	1.72	7.48	78.79
T5 base	Soft domain shift	9.21	25.71	86.44	7.28	24.13	86.03
Flan-T5 base	Zero-shot (task shift)	4.08	15.32	87.00	3.71	17.25	86.44
Flan-T5 base	In-distribution	8.80	23.19	87.49	6.15	20.84	86.48
Flan-T5 base	Hard domain shift	6.89	20.53	87.16	4.32	17.00	85.88
Flan-T5 base	Soft domain shift	10.38	27.17	88.22	7.18	23.04	86.90
Flan-T5 large	Soft domain shift	14.37	33.74	88.21	10.90	30.05	87.44
T5 XL	Zero-shot (task shift)	2.05	8.28	81.90	2.28	9.73	80.37
T5 XL	Soft domain shift	34.14	53.55	91.40	18.82	38.26	88.81
Flan-T5 XL	Zero-shot (task shift)	2.70	12.72	86.72	2.88	16.20	86.52
Flan-T5 XL	In-distribution	11.49	28.96	88.90	16.61	36.27	89.40
Flan-T5 XL	Hard domain shift	29.55	48.17	91.39	8.37	25.06	87.56
Flan-T5 XL	Soft domain shift	32.81	52.21	92.16	18.69	38.72	89.75

Table 12: Results of the definition generation experiments.

C Additional Results

C.1 Zero-Shot Evaluation of Flan-T5 (Task Shift)

Here we directly evaluate Flan-T5 XL on the WordNet and Oxford test sets, without any fine-tuning nor in-context learning.¹⁴ Table 3 in the main paper shows low BLEU and ROUGE-L scores but rather high BERT-F1. Overall, the model does not exhibit consistent task understanding (e.g. it generates ‘SKEPTICISM’ as a definition for ‘healthy’ as exemplified in the phrase ‘healthy skepticism’). A qualitative inspection, however, reveals that the generated definitions can still be often informative (e.g., ‘A WORKWEEK THAT IS LONGER THAN THE REGULAR WORKWEEK’ is informative with respect to the meaning of ‘overtime’ although the ground truth definition is ‘BEYOND THE REGULAR TIME’). The two surface-overlap metrics cannot capture this, but the relatively high BERT-F1 confirms that the semantic content of generations is largely appropriate. There are indeed also many good zero-shot definitions. For example ‘INTENSE’ for ‘fervent’ as in ‘the fervent heat’, or ‘A CONVERSATION’ for ‘discussion’ in ‘we had a good discussion’.

C.2 Other Models and Model Variants

We evaluate T5 (base and XL) and Flan-T5 (base, large, and XL) under the same generalisation conditions presented for Flan T5 XL in the main paper

¹⁴We only condition generation on the usage examples and the task prompt. We do *not* provide full instances (i.e., usage examples, task prompts, and definitions) in the context, as one would do in a few-shot setup.

(Section 4.1) and above in Appendix C.1. Results for FlanT5-XL are reported in the main paper (Table 3); here, in Table 12, we report results for all models and model variants.

C.3 Evaluation Cards

In Table 13, we provide an evaluation card to clarify the nature of the generalisation tests performed on definition generators.¹⁵ In-distribution tests are not included as they do not include any shift between the training and test data distributions (Hupkes et al., 2022). We also register our work in the GenBench evolving survey of generalisation in NLP.¹⁶

D Additional Examples of Generated Definitions and Sense Labels

Some definitions generated by Flan-T5 XL manage to capture very subtle aspects of the contextual lexical meaning. In the following list, we give the usage and then the contextual definition of ‘word’:

1. ‘There are people out there who have never heard of the Father, Son and Holy Spirit, let alone the **Word** of God.’: ‘THE BIBLE’
2. ‘Good News Bible Before the world was created, the **Word** already existed; he was with God, and he was the same as God.’: ‘(CHRISTIANITY) THE SECOND PERSON OF THE TRINITY ; JE’
3. ‘It was in that basement that I learned the skills necessary to succeed in the difficult thespian world-specifically, get up on stage, say

¹⁵https://genbench.org/eval_cards

¹⁶<https://genbench.org/references>

Motivation					
<i>Practical</i>	<i>Cognitive</i>	<i>Intrinsic</i>	<i>Fairness</i>		
□ △ ○					
Generalisation type					
<i>Compositional</i>	<i>Structural</i>	<i>Cross Task</i>	<i>Cross Language</i>	<i>Cross Domain</i>	<i>Robustness</i>
		□		△ ○	
Shift type					
<i>Covariate</i>	<i>Label</i>	<i>Full</i>		<i>Assumed</i>	
△ ○		□			
Shift source					
<i>Naturally occurring</i>	<i>Partitioned natural</i>	<i>Generated shift</i>		<i>Fully generated</i>	
□ △ ○					
Shift locus					
<i>Train-test</i>	<i>Finetune train-test</i>	<i>Pretrain-train</i>	<i>Pretrain-test</i>		
	△ ○		□		

Table 13: Evaluation card for the generalisation tests performed on definition generators. The setups are: zero-shot (□), hard domain shift (△), and soft domain shift (○). In-distribution tests are not included as they do not include any shift between the training and test data distributions.

my words, get off the stage-skills...’: ‘THE DIALOGUE OF A PLAY.’

Interesting insights can be drawn from how the embeddings of the generated definitions are located in the vector space. Figure 8 shows PCA projections of definition embeddings for usages of the words ‘*chef*’ and ‘*lass*’ from the English DWUG. Colours represent sense clusters provided in the DWUG, and the legend shows most prototypical definitions for each sense generated by our best system (singleton clusters are ignored). The large star for each sense corresponds to its sense label (as opposed to smaller stars corresponding to other definitions not chosen as the label).

For the word ‘*chef*’, there are two sense clusters, for which an identical definition is chosen (‘A COMMANDER’). This most probably means that these clusters should in fact be merged together, or that they are in the process of splitting (see also Section 7). These two senses are (not surprisingly) much closer to each other than to the definitions from the ‘PROFESSIONAL COOK’ sense. For the word ‘*lass*’, it is interesting how separate is a small bluish group of definitions in the bottom right corner of the plot, where the target form is actually ‘*lassi*’. The fine-tuned Flan-T5-XL model defined this group as ‘A COLD DRINK MADE FROM MILK

CURDLED BY YOGURT’, which is indeed what ‘*lassi*’ is (ignoring minor details).

E Human Evaluation Guidelines

Figures 9 and 10 show the results of the human evaluation.

‘You are given a spreadsheet with four columns: **Targets**, **Examples**, **System1** and **System2**. In every row, we have one target English word in the **Targets** column and five (or less) example usages of this word in the Examples column. Usages are simply sentences with at least one occurrence of the target word: one usage per line.

Every row is supposed to contain usages where the target word is used in the same sense: this means that for ambiguous words, there will be multiple rows, each corresponding to a particular sense. This division into senses is not always 100% correct, but for the purposes of this annotation effort, we take it for granted. Note that the five example usages in each row are sampled randomly from a larger set of usages belonging to this sense.

System1 and System2 are computational models which produce human-readable labels or definitions for each sense of a target word. They employ different approaches, and your task is to compare and evaluate the labels generated by these two systems. Note that in each row, the names ‘System1’ and ‘System2’ are randomly assigned to the actual generation systems.

The generated sense labels are supposed to be useful for historical linguists and lexicographers. Thus, they must be:

1. **Truthful:** i.e., should reflect exactly the sense in which the target word is occurring in the example usages. Ideally, the label should be general enough to encompass all the usages from the current row, but also specific enough so as not to mix with other senses (for poly-semantic target words).
2. **Fluent:** i.e., feeling like natural English sentence or sentences, without grammar errors, utterances broken mid-word, etc

You have to fill in the **Judgements** column with one of six integer values:

- **0:** both systems are equally bad for this sense
- **1:** System 1 is better, but System 2 is also OK
- **11:** System 1 is better, and System 2 is bad

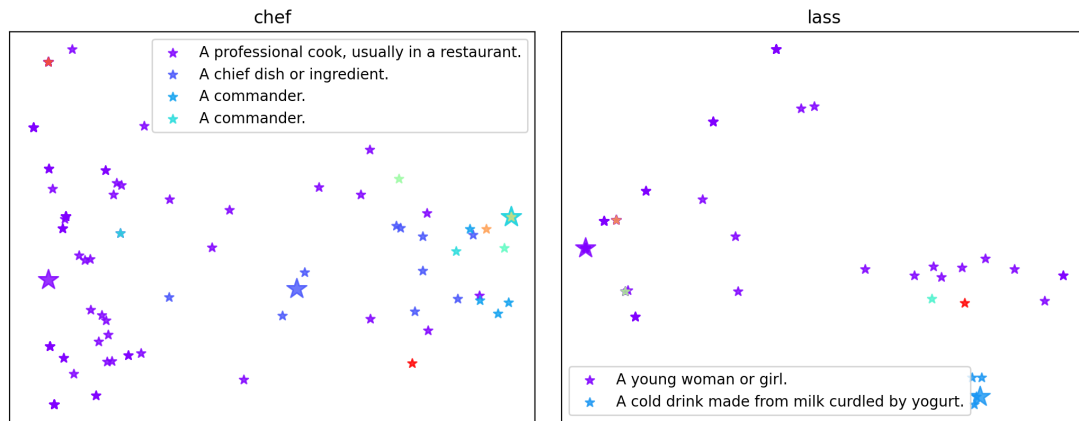


Figure 8: PCA projections of definition embeddings for two target words from English DWUG.

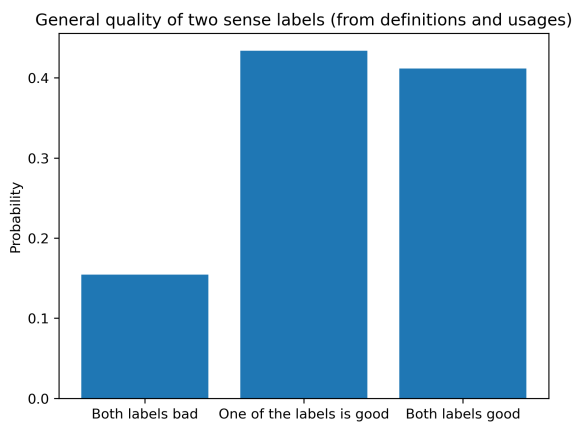


Figure 9: General quality of generated sense labels

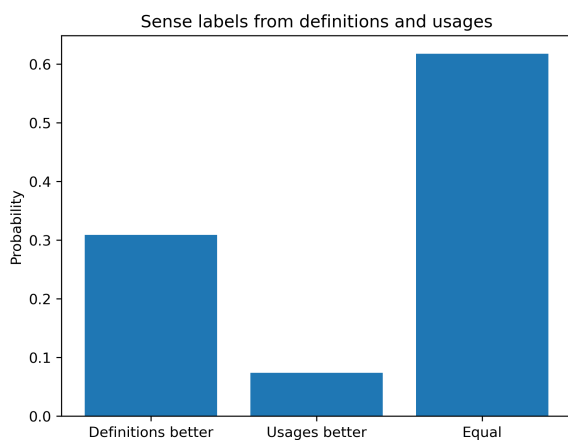


Figure 10: Human comparison of sense labels produced from definitions and from usages

- **2**: System 2 is better, but System 1 is also OK
- **22**: System 2 is better, and System 1 is bad
- **3**: both systems are equally good for this sense

Some rows are already pre-populated with the **3** judgement, because the sense labels generated by both systems are identical. We hypothesise that this most probably means that both labels are equally good. Please still have a look at these identical labels and change **3** to **0** in case you feel that in fact they are equally bad.'

F Sense Dynamics Maps

It is easy to find different sense clusters which are assigned *identical* definition labels. Usage examples from sense clusters c_2 and c_3 for the word 'chef', to which our system assigned the same label: 'A COMMANDER':

- c_2 : 'He boasted of having been a **chef** de brigade in the republican armies of France', 'Morrel has received a regiment, and Joliette is **Chef** d'Escadron of Spahis', 'as major-general and **chef** d'escadron, during the pleasure of our glorious monarch Louis le Grand'
- c_3 : 'That brave general added to his rank of **chef** de brigade that of adjutant general', 'I frequently saw Mehevi and several other **chefs** and warriors of note take part'

Thus, a user can safely accept the suggestion of our system to consider these two clusters as one sense.

Note that 'A COMMANDER' practically disappeared as a word sense in the 20th century, replaced by 'A PROFESSIONAL COOK, USUALLY IN A RESTAURANT'.

G Clustering Embedding Spaces

We constructed three types of embedding spaces; (i) contextualised token embeddings, (ii) sentence embeddings, and (ii) definition embeddings. We did so for two language models: RoBERTa-large and DistilRoBERTa. Since we cluster the embedding spaces for each target word individually, we obtain different optimal number of clusters for each target word. Table 5 displays the average results over all target words.

We observe that the optimal number of clusters K is substantially higher for the definition embedding spaces for both RoBERTa-large and DistilRoBERTa. However, this is an artefact of the data: since some distinct usages yield identical definitions for a target word, the definition space often-times consist of less distinct data points, which greatly impacts the average silhouette scores. Future work should point out what clustering methods are most applicable to definition embedding spaces. Still, this decrease in data points confirms how the definition embedding space could represent usages at a higher level of abstraction, collapsing distinct usages into identical representations.

Figure 11 displays the T-SNE projections of each of the three embedding spaces of RoBERTa-large. As for Distil-RoBERTa, the definition embedding space appears to have spacial properties that are more similar to contextualised *token* embedding spaces than to *sentence* embedding spaces: the definition embeddings are more separated than the sentence embeddings, and are cluttered in a similar manner as the token embeddings.

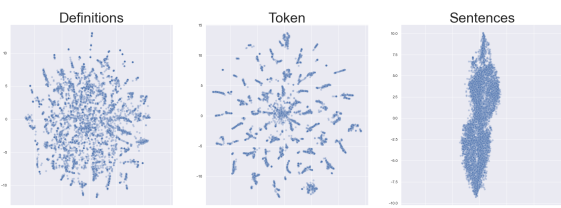


Figure 11: T-SNE projection of each embedding space, RoBERTa-Large model.

Model	Representation	Inter-cluster	Intra-cluster	Ratio
RoBERTa-large	Sentence	0.017	0.013	1.248
	Token	0.042	0.034	1.272
	Definitions	0.008	0.006	1.349
DistilRoBERTa	Sentence	0.665	0.592	1.126
	Token	0.591	0.477	1.258
	Definitions	0.705	0.509	1.397

Table 14: Average Separation and Cohesion scores of each cluster for each target word from the English DWUG

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Unnumbered "Limitations" section after the Conclusion (section 8)
- A2. Did you discuss any potential risks of your work?
Unnumbered "Limitations" section after the Conclusion (section 8)
- A3. Do the abstract and introduction summarize the paper's main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3

- B1. Did you cite the creators of artifacts you used?
3
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The licenses are described in the papers we cite.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We did not collect or use any such data.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
This is described in the papers we cite.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C Did you run computational experiments?

4,5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4 and Appendix
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4,5
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
6
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
6
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The human annotator co-authored the paper, so this discussion was not necessary.
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We did not collect any data.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
6