SemEval-2020 Task 3: Graded Word Similarity in Context

Carlos S. Armendariz[*] and **Matthew Purver**[*†]
[*]Cognitive Science Research Group, Queen Mary University of London, London, UK
{c.santosarmendariz, m.purver}@qmul.ac.uk

Senja Pollak[†] and **Nikola Ljubešić**[†]
[†]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{senja.pollak, nikola.ljubesic}@ijs.si

Matej Ulčar and **Marko Robnik-Šikonja**
University of Ljubljana, Faculty of Computer and Information Science, Slovenia
{matej.ulcar, marko.robnik}@fri.uni-lj.si

Ivan Vulić
University of Cambridge, UK
iv250@cam.ac.uk

Mohammed Taher Pilehvar
Tehran Institute for Advanced Studies
mp792@cam.ac.uk

Abstract

This paper presents the *Graded Word Similarity in Context (GWSC)* task which asked participants to predict the effects of context on human perception of similarity in English, Croatian, Slovene and Finnish. We received 15 submissions and 11 system description papers. A new dataset (CoSimLex) was created for evaluation in this task: it contains pairs of words, each annotated within two short text passages. Systems beat the baselines by significant margins, but few did well in more than one language or subtask. Almost every system employed a Transformer model, but with many variations in the details: WordNet sense embeddings, translation of contexts, TF-IDF weightings, and the automatic creation of datasets for fine-tuning were all used to good effect.

1 Introduction

Contextualised word embeddings, produced by models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), have quickly become the standard in NLP systems. They deliver impressive performance in language modeling and downstream tasks; but there are few resources available which allow intrinsic evaluation in terms of the properties of the embeddings themselves, or their ability to model human perception of meaning, and how these depend on context. For non-contextualised models, resources like WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) were instrumental to evaluate their ability to reflect human similarity judgements. However these datasets treat pairs of words in isolation, and thus cannot tell us much about the effect of context. The few resources that work with context, like SCWS (Huang et al., 2012), WiC (Pilehvar and Camacho-Collados, 2019), and WSim (Erk et al., 2013), focus on word sense and discrete effects, thus missing the more graded effects that context has on words in general, and that approaches like ELMo and BERT would seem well suited to model. Further, USim (Erk et al., 2013) focuses on separate sentential contexts only in the English language.

The goal of SemEval-2020 Task 3: Graded Word Similarity in Context, was to move towards filling that gap. We created a new dataset, **CoSimLex** (Armendariz et al., 2020), which builds on the familiar pairwise, graded similarity task of SimLex-999, but extends it to pairs of words as they occur in context; specifically, each pair of words appears together in two different shared contexts (see Figure 1). The task was designed to test the ability of participating systems to reflect human judgements of word meaning similarity in context, and crucially, the way in which this varies as context is changed. In addition, since CoSimLex takes the *gradedness* of human judgements into account, the task applies not only to polysemous words, or words with distinct senses, but to the phenomenon of context-dependency of word meaning in general. The dataset is also multi-lingual: besides English, it includes three less-resourced European languages, Croatian, Finnish, and Slovene.

Word1: man **Word2: warrior**	**SimLex**: μ 4.72 σ 1.03
Context1	**Context1:** μ 7.88 σ 2.07
When Jaimal died in the war, Patta Sisodia took the command, but he too died in the battle. These young **men** displayed true Rajput chivalry. Akbar was so impressed with the bravery of these two **warriors** that he commissioned a statue of Jaimal and Patta riding on elephants at the gates of the Agra fort.	
Context2	**Context2:** μ 3.27 σ 2.87
She has a dark past when her whole family was massacred, leaving her an orphan. By day, Shi Yeon is an employee at a natural history museum. By night, she's a top-ranking woman **warrior** in the Nine-Tailed Fox clan, charged with preserving the delicate balance between **man** and fox.	
	P-Value: 1.3×10^{-6}

Figure 1: Example from the English dataset, showing a word pair with two contexts, each with mean and standard deviation of human similarity judgements. The original SimLex values for the same word pair without context are shown for comparison. The P-Value shown is the result of a Mann-Whitney U test.

2 Background

Our motivation lies in the cognitive and psychological mechanisms by which context affects our perception of word meaning. Here, we present two of the most prominent ideas that helped define the task and dataset, and explain why previous datasets for similarity in context are not well suited to test them.

2.1 Contextual Modulation

One debate in lexical semantics is whether the discreteness of lexical senses is fundamental or just a perception. Cruse (1986) proposed a compromise, distinguishing two different manners in which sentential context modifies the meaning of a word. First, the context can select for different discrete senses; in this case, the word is described as *ambiguous*, and the process as **contextual selection of senses** (familiar from many word sense disambiguation tasks). Second, the context can modify meaning within the scope of a single sense by *highlighting* certain semantic traits and *backgrounding* others. This is described as **contextual modulation of meaning**, and the word as *general* with respect to the traits being modulated. This latter effect is not discrete, but continuous or graded; every word is *general* to some extent, and thus has a different meaning in every context in which it appears.

1. At this point, the bank was covered with brambles.
2. Sue is visiting her pregnant cousin.
3. Arthur poured the butter into a dish.

The main effect of the context in example (1) is to *select* one of the discrete senses associated with the word *bank*. In contrast, in examples (2) and (3), the contexts *modulate* the meanings of the words *cousin* and *butter*: for *cousin*, promoting the "female" trait, and for *butter*, the "liquid" trait. This is possible because of the *general* quality of these words. Other traits could be promoted in different contexts: *cousin* includes male and female, but also tall, short, happy and sad cousins. Related traits can be promoted as a consequence of this modulation: we understand the butter as not only liquid, but warm. We expect this to affect similarity judgements.

2.2 Salience Manipulation

In contrast to this purely linguistic view, we can take a cognitive perspective on language and meaning, seeing it as a more general expression of human cognition (Evans and Green, 2018). In this view, the units of interest are the *conceptual structures* associated with words or lexical units, rather than the words themselves. One approach is to see these in terms of *conceptual spaces* characterised by *quality dimensions* (Gärdenfors, 2000; Gärdenfors, 2014). These dimensions may be concrete (weight, temperature, brightness) or abstract (awkwardness, goodness), and concepts are defined as regions (usually convex) within the space. This space is not fixed: when we communicate we constantly re-negotiate the dimensions framing the conversation and their salience (Warglien and Gärdenfors, 2015). This **salience manipulation** changes their perceived importance. Priming effects are proposed as the main mechanism that facilitates this process (Pickering and Garrod, 2004). This type of semantic effect was first reported by Meyer and Schvaneveldt (1971) when they found that their lexical decision task was responded to faster when the subjects were primed with words associated to the target words.

From this perspective, then, context affects meaning not via the presence of specific words, but via a change in the *mental state* of the hearer/reader.

1. My muffins were a failure, I should have used butter or margarine instead of olive oil.
2. Vegan chefs replace animal fats, like butter, with plant based ones like olive oil or margarine.
3. Vegan influencers believe the consumption of animal products is cruel and unnecessary.

In example (1), the context of baking increases the salience of dimensions related to physical properties of ingredients; butter and margarine (both solid) therefore seem more similar to each other than to olive oil (liquid). In contrast, example (2)'s context of veganism makes the animal vs. plant-based dimension very salient; margarine and olive oil now seem more similar to each other than to the animal-based butter.

The effects of *salience manipulation* and *contextual modulation* have important differences. The effect in example (3) is introduced by the word *poured* and limited to the word *butter*, but the effect in example (1) seems more general: once a context triggers changes in the salience of conceptual dimensions, any word thereafter is affected. Our hypothesis is that the *salience manipulation* effect applies even when the target words are not present: a context like example (3) will impact later perceptions of similarity of butter, margarine and olive oil. We hope to test such predictions in later analyses.

2.3 Related Work

The **Stanford Contextual Word Similarity (SCWS)** dataset (Huang et al., 2012), and the similar **USim** dataset (Erk et al., 2013) contain graded similarity judgements of pairs of words in the context of naturally occurring sentences (e.g., from Wikipedia with SCWS). However, the datasets were designed to evaluate a discrete multi-prototype model, so the focus was on contexts that select for discrete word senses, and each word in a pair was presented in its own distinct context. This prevents a systematic comparison of contextual effects on pairwise similarity. In addition, inter-rater agreement (IRA) on SCWS, measured as the Spearman correlation between different annotators, shows worryingly low scores. As Pilehvar and Camacho-Collados (2019) point out, the mean IRA between each annotator and the average of the rest, considered a human-level upper bound for model performance, is 0.52; while the performance of a simple context-independent model like word2vec (Mikolov et al., 2013) is 0.65. Many scores also show a very large standard deviation, with annotators rating the same pair very differently. One possible reason may lie in the annotation design: the task itself does not directly enforce engagement with the context, and the target words were presented to annotators highlighted in boldface, making it easy to pick them out from the context without reading it.

Some of these limitations were addressed by the more recent **Words-in-Context (WiC)** dataset (Pilehvar and Camacho-Collados, 2019). With a more direct and straightforward take on word sense disambiguation, each entry of the dataset is made of two lexicographer examples of the same word, and labelled as to whether the word sense in the two examples/contexts is the same or different. This forces engagement with the context; it also creates a task in which context-independent models like word2vec "would perform no better than a random baseline"; and inter-rater agreement scores are much more healthy. However, as the dataset focuses on discrete word senses, it cannot capture graded effects of context.

These datasets are also available only in English. Multi-lingual similarity datasets exist: in **SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity**, Camacho-Collados et al. (2017) used five different languages, and even used pairs in which each word was presented in a different language. A more recent **Multi-SimLex** dataset (Vulić et al., 2020) comprises similarity ratings for 1,888 concept pairs aligned across 13 typologically diverse languages. However, the pairs in both datasets were annotated out of context, preventing analysis of contextual effects.

3 Task Description

Our dataset is based on pairs of words from SimLex-999 (Hill et al., 2015). Each instance is a naturally-occurring context, taken from Wikipedia, in which both words in the pair appear, labelled with a similarity score given by human annotators. For each pair, the dataset contains two different contexts (see Section 4 for more detail on dataset and choice of contexts). We proposed two different subtasks: first, to predict the change in similarity score between the two different contexts for each pair; second, to predict the similarity scores themselves. These are related but independent tasks that use the same input data, but each subtask had its own phases and leaderboards. Submissions for each subtask were independent and participants were able to use different models for each subtasks and each language. The tasks were unsupervised, and so no training data was released; However, we released a small *practice kit* which contained a practice dataset, a script to generate the baseline and evaluation scripts so participants could easily reproduce results, and understand how the dataset looked and how the task was evaluated.

3.1 Subtask 1: Predicting Change

In the first subtask, participants were asked to predict the *change* in the similarity ratings of a pair of words when the human annotators are presented with the same word pair within two different contexts. This task directly addresses our main question. It evaluates how well systems are able to model the effect that context has in human perception of similarity. Theoretically a model could perform very well at modelling change without actually being able to accurately predict the ratings themselves. On the other hand, any context-independent model will predict no change and perform poorly in this task.

3.2 Subtask 2: Predicting Contextual Ratings

In the second subtask, participants were asked to predict the absolute similarity rating for each pair in each context. This is a more traditional task which evaluates systems' ability to model both similarity of words and the effect that context has on it. Good context-independent models could theoretically give reasonably competitive results in this task, however we still expect context-dependent models to have a considerable advantage.

4 Dataset

CoSimLex (Armendariz et al., 2020) is based on pairs of words from SimLex-999 (Hill et al., 2015); the reliability and common use of SimLex makes it a good starting point and allows comparison of judgements and model outputs to the context-independent case. For Croatian and Finnish we use existing translations of SimLex-999 (Mrkšić et al., 2017; Venekoski and Vankka, 2017; Kittask, 2019). In the case of Slovene, we have produced our own new translation,[1] following Mrkšić et al. (2017)'s methodology for Croatian.

The dataset consists of 340 pairs in English, 112 in Croatian, 111 in Slovene and 24 in Finnish. Each pair is rated within two different contexts, giving a total of 1174 scores of contextual similarity. This poses a difficult task: to find suitable, organically occurring contexts; this task is even more challenging for languages with less resources, and as a result the selection of pairs is different for each language.

Each line of CoSimLex is made of a pair of words selected from SimLex-999; two different contexts extracted from Wikipedia in which these two words appear; two scores of similarity, each one related to one of the contexts, calculated as the mean of annotator ratings for that context; two scores of standard deviation; the p-value given by applying the Mann-Whitney U test to the two score distributions; and the four inflected forms of the words exactly as they appear in the contexts (including case; note that in the morphologically rich languages, many inflections are possible). To the best of our knowledge, this is the first reasonably sized dataset in which differences in contextual similarity between two words are supported with a test of statistical significance. Figure 1 shows an example from the English dataset.

4.1 Context Selection

For each word pair we needed to find two suitable contexts. These contexts were extracted from each language's Wikipedia. They are made of three consecutive sentences and they needed to contain the pair

[1] Available from `http://hdl.handle.net/11356/1309`

of words, appearing only once each. English is by far the easiest language to work with, not only because of the amount and quality of the text contained in the English version of Wikipedia but because the other three languages are highly inflected (Croatian, Finnish and Slovene). To overcome this, we worked with data from (Ginter et al., 2017)[2] which contains tokenised and lemmatised versions of Wikipedia for 45 languages.

The differences were expected to be small; to maximise the chance of finding contexts that produced different ratings of similarity, we used a dual process based on ELMo and BERT models. First, we used a model to rate the similarity between the target words within each of the candidate contexts; then selected the context in which it scored the pair as the most similar, and the context in which it scored them as most different. We repeated the process using both ELMo and BERT scores. This gave us 4 promising contexts. Then we added 4 randomly selected contexts for a total of 8 candidate contexts.

The final selection of two contexts was made by expert human annotators, one per language. Our experts were presented with 8 candidate contexts and asked to select the two that maximised the potential contrast in similarity. In the case of less-resourced languages, the smaller size and lower quality of the Wikipedia text resources required some extra steps to ensure the quality of the final annotation. A set of heuristic filters were used to try to remove badly constructed contexts. In addition we produce 16 candidates instead of 8 for the expert annotators to choose from.

4.2 Annotation

As starting point for our annotation methodology, we adapted the instructions used for SimLex-999. This way we benefited from its tested method of explaining how to focus on *similarity* rather than *relatedness* or *association* (Hill et al., 2015). As explained in their original paper, *cup* and *mug* are very similar, while *coffee* and *cup* are strongly related but not similar at all. For English we adopted their crowd-sourcing process: we used *Amazon Mechanical Turk*, with the same initial scoring scale (0 to 6), which is later transformed to a 0 to 10 scale. For the less-resourced languages, crowdsourcing is not a viable option due to lack of available speakers, and we recruited annotators directly. This means fewer annotators (for Croatian, Finnish and Slovene, 12 annotators vs 27 in English), however the average quality of annotation is higher and the data requires less post-processing.

In regards to the annotation process itself, our goal is to capture the kind of contextual phenomena discussed in Section 2: lexical meaning modulation and conceptual salience manipulation. In order to maximise our chances we defined three goals:

- Interaction with the context should be as natural as possible, so as to maximise priming effects and capture the potential change in the salience of conceptual dimensions.
- Annotators should have the chance to account for lexical modulation within the sentence.
- The process should ensure that the annotators engage fully with the context.

With these goals in mind we designed a two-step mixed annotation process. Our online survey interface is composed of two pages per pair of words and context (each annotator scores only one of the contexts). In the first page the annotators are presented with the context, and asked to read it and come up with two words “inspired by it”. Once this is complete, the second page shown presents the context again, but with the target words now highlighted in bold; they are now asked to rate the similarity of target words within the sentence. Notice these target words are completely independent to the ones that were chosen as “inspired by the context” (see Apendix A for an example of the survey).

The second page is the main scoring task; it is designed to capture changes in scores of similarity due both to lexical modulation and — because we hope the annotators are still primed by their recent previous engagement with the context — the changes in the salience of conceptual dimensions. The separate task on the first page is intended to make annotators engage fully with the whole context, while maintaining a natural interaction with it to maximise any priming effects. One of the possible problems we identified in the previous SCWS annotation process is the fact that the words were always highlighted in bold, making it easy for annotators (Amazon Mechanical Turk workers) to just look at the pair of words in isolation and

[2] Available from `http://hdl.handle.net/11234/1-1989`

Dataset	#pairs	Sim	StDev	Spearman's ρ	Change (Abs)	$p < 0.1$	$p < 0.05$
SimLex-999	999	4.56	1.27	0.78	-	-	-
English CoSimLex	340	5.54	2.24	0.77	2.16	65%	61%
Croatian CoSimLex	112	4.39	2.23	0.76	2.32	65%	54%
Slovene CoSimLex	111	4.90	2.17	0.77	1.96	59%	46%
Finnish CoSimLex	24	4.08	2.16	0.81	1.75	33%	29%

Table 1: Similarity, standard deviation, Spearman's ρ and change are average values. The two rightmost columns denote the proportion of pairs whose differences of scores with the original values are statistically significant at p-value < 0.1 and p-value < 0.05.

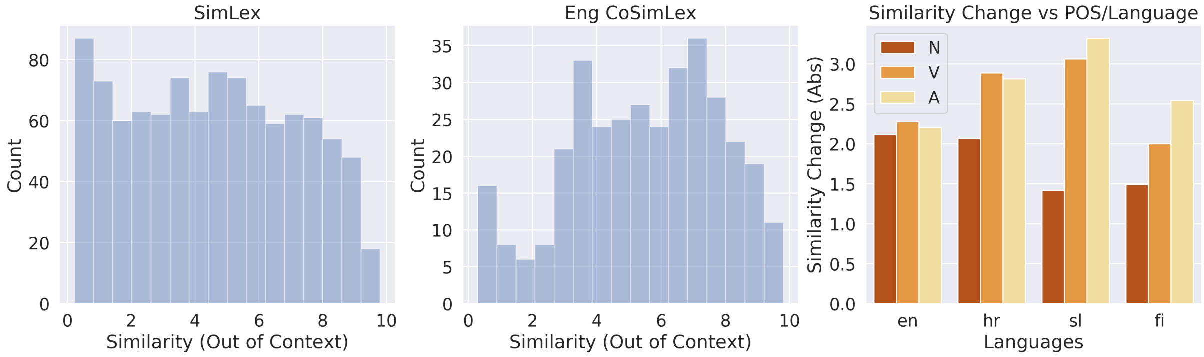


Figure 2: **(a) (b):** Differences in the distribution of similarity between SimLex-999 and the English CoSimLex; **(c):** Change in the scoring of similarity between contexts categorized by language and part of speech

to not read the rest of the contexts. Our initial task is designed to prevent this (the words are not in bold in the first page).

4.3 Post-Processing

Post-processing and cleaning the data is especially important when relying on crowd-sourcing platforms to source annotators. Reliability of annotation was ensured by an adapted version of SimLex-999's post-processing method, which includes rating calibration and the filtering of annotators with very low correlation to the rest, see the original paper for details (Hill et al., 2015). In addition, we were able to use responses to the first annotation question to check annotator engagement with the context.

In English there were instances in which a block of annotations resulted in especially bad data. In those cases the only solution was repeating the annotation of the whole block. In our experience, obtaining good annotation using Amazon Mechanical Turk is not straightforward, but can be improved by a few strategies to attract good annotators. It is possible to engage with quality annotators and create private tasks for them inside the platform, which produces better data and allows higher payment for the worker. We encourage other researchers to use similar strategies when possible. This was not an issue with the rest of the languages, where annotators were sourced directly. After the post-processing steps the English dataset retained an average of 21 annotations per entry (from a starting point of 27) while the rest of the languages kept an average of 10 annotations (from the starting 12).

4.4 Basic Analysis

The difficulty of finding contexts for the less-resourced languages restricted the selection of pairs available. As a consequence the overlap of pairs between different languages is smaller than originally intended (86 pairs appear in two languages, 12 in three and only 4 appear in all languages). However we were still able to replicate SimLex-999's proportions of nouns, verbs and adjectives (about two thirds nouns, two ninths verbs and one ninth adjectives). In English we checked other metrics, namely concreteness, standard deviation and out-of-context similarity. The first were kept in similar ranges to SimLex, however for out-of-context similarity we decided to lower the proportion of antonyms and low similarity score pairs, which as noted by Camacho-Collados et al. (2017) were substantially overrepresented (see Figure 2).

We expected that the relative complexity of the annotation process and the increased confounding effects could affect inter-rater agreement; however, as we can see in Table 1, the different CoSimLex datasets show correlation scores very close to SimLex-999's IRA (ρ = 0.77 vs ρ = 0.78 in English). In the same table we can see the standard deviation is higher. Differences in the average similarity score are mainly due to the pair selection. After the post-processing and cleaning of the data both the crowdsourced and directly sourced annotation produced similar IRA and standard deviation. We wondered if the highly inflected nature of some of the languages might increase the contextual effects; but as can be seen in the table, the average change is very similar, even lower for Slovene and Finnish. However an interesting phenomenon seems to appear when we look at the distribution by part of speech; Chart (c) in Figure 2 suggest that verbs and adjectives in Croatian, Slovene and Finnish do see an increased effect of context compared with English ones. Importantly, the global percentage of statistically significant results is high (indeed, higher than we expected), with a global 62% of pairs showing statistically significant differences between contexts.

One potential confounding effect is the separation between words as presented in context (the number of intervening words between the target pair): it is possible this could affect annotators' perception of similarity. There is a very small negative correlation between similarity ratings and distance (Pearson r = -0.13). The source of this could be annotator bias, a linguistic effect or a combination of the two; but the effect seems small enough to ignore for current purposes.

5 Evaluation Metrics

The first subtask looked at the change in similarity between the two contexts, therefore it was important to preserve the difference between positive and negative values since it reflected in which of the two context the system believed the two words to be more or less similar. Consequently the most appropriate metric was **Uncentered Pearson Correlation** which looks at the deviation from zero instead of the mean.

$$CC_{uncentered} = \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sqrt{(\sum_{i=1}^{n} x_i)^2 (\sum_{i=1}^{n} y_i)^2}}$$

For the second subtask, which looked at the more traditional absolute value of similarity in context, we followed (Camacho-Collados et al., 2017) and used the harmonic mean of the Pearson and the Spearman correlations between the system's results and the average of the human annotations.

6 Baselines

Our task studies contextual effects in four different languages, which made Multiligual BERT the perfect candidate for our baseline. Released shortly after the original BERT model (Devlin et al., 2019), it employs its same architecture while being trained in more than 100 different languages, our four languages between them. The original model introduced an innovative masking strategy that for the first time allowed for a bidirectional Transformer language model. BERT models are renowned for their ability to capture contextual effects, ability which is often blamed for an important part of their performance improvements. For the baseline of our task we used the uncased version of the model, and as a common strategy we used the contents of the last layer to form our embeddings. BERT creates sub-word tokens for the out of vocabulary words, in those cases our strategy was simply averaging the sub-word vectors to form a word embeddings.

Additionally, the results achieved by ELMo are added to Tables 2 and 3 as a reference. This model precedes BERT and was one of the first to produce contextualised embeddings (Peters et al., 2018), in this case using a bidirectional LSTM. The original ELMo dataset was only trained in English, however we used ELMo models recently trained in Croatian, Slovene and Finnish (Ulčar and Robnik-Šikonja, 2020).

7 Participants & Results

The task received a total of 14 submissions for the first subtask and 15 submissions for the second. From those, 11 teams submitted system description papers for review. In order to be considered for the official rankings we asked participants to fill a form with some basic information about their systems. Teams that

SUBTASK 1							
English		Croatian		Slovene		Finnish	
Ferryman	0.774	BabelEnconding	0.74	Hitachi	0.654	will_go	0.772
will_go	0.768	Hitachi	0.681	BRUMS	0.648	Ferryman	0.745
MultiSem	0.76	BRUMS	0.664	BabelEnconding	0.646	BabelEnconding	0.726
LMMS	0.754	Ferryman	0.634	CiTIUS-NLP	0.624	BRUMS	0.671
BRUMS	0.754	LMMS	0.616	Ferryman	0.606	CiTIUS-NLP	0.671
Hitachi	0.749	will_go	0.597	will_go	0.603	MultiSem	0.593
BabelEnconding	0.73	CiTIUS-NLP	0.587	LMMS	0.56	Hitachi	0.574
CiTIUS-NLP	0.721	MineriaUNAM	0.374	MineriaUNAM	0.328	MineriaUNAM	0.389
MineriaUNAM	0.544	MultiSem	-	MultiSem	-	LMMS	0.36
JUSTMasters	0.738		0.44		0.512		0.546
UZH	0.765		-		-		-
mBERT_uncased	0.713		0.587		0.603		0.671
ELMo	0.570		0.662		0.452		0.550

Table 2: Subtask 1 Final Ranking: The values are calculated as the Pearson Uncentered Correlation between the system's scores and the average human annotation. It represents the system's ability to predict the change in perception produced by the contexts. Since different annotators looked at each context, human performance couldn't be calculated for this subtask. JUSTMasters and UZH are not part of the official ranking since they were able to optimise their systems with more than the competition's limit of 9 submissions.

SUBTASK 2							
English		Croatian		Slovene		Finnish	
MineriaUNAM	0.723	BabelEnconding	0.658	BabelEnconding	0.579	BRUMS	0.645
LMMS	0.72	Hitachi	0.616	BRUMS	0.573	BabelEnconding	0.611
AlexU-Aux-Bert	0.719	MineriaUNAM	0.613	CiTIUS-NLP	0.538	MineriaUNAM	0.597
MultiSem	0.718	LMMS	0.565	will_go	0.516	MultiSem	0.492
BRUMS	0.715	BRUMS	0.545	AlexU-Aux-Bert	0.516	Ferryman	0.357
will_go	0.695	CiTIUS-NLP	0.496	Hitachi	0.514	LMMS	0.354
Hitachi	0.695	AlexU-Aux-Bert	0.402	MineriaUNAM	0.487	will_go	0.35
CiTIUS-NLP	0.687	will_go	0.402	LMMS	0.483	Hitachi	0.335
BabelEnconding	0.634	Ferryman	0.397	Ferryman	0.345	CiTIUS-NLP	0.289
Ferryman	0.437	MultiSem	-	MultiSem	-	AlexU-Aux-Bert	0.289
JUSTMasters	0.725		0.443		0.44		0.68
mBERT_uncased	0.573		0.402		0.516		0.289
ELMo	0.510		0.529		0.407		0.516
Human	0.77		0.76		0.77		0.81

Table 3: Subtask 2 Final Ranking: The values are calculated as the harmonic mean of the Spearman and Pearson correlation between the system's scores and the average human annotation. It represents the system's ability to predict contextual human perception of similarity. Human performance is the average value when comparing each annotator against the average of the rest. JUSTMasters is not part of the official ranking since they were able to optimise their system with more than the competition's limit of 9 submissions.

neither filled the form nor submitted a system description paper do not appear in the official rankings (Tables 2 and 3). We will discuss here the results of the remaining 11 systems.

First, we describe a group of systems designed around sense embeddings created using WordNet (Miller, 1995) as a guide. The most successful was the submission by **LMMS**. They employed a similar strategy to the one set out in (Loureiro and Jorge, 2019), creating pretrained embeddings for each sense in WordNet, this time using XLM-R (Conneau et al., 2019) and SemCor augmented with their own UWA dataset (Loureiro and Camacho-Collados, 2020). This approach achieved second place in the English Subtask 1 and fourth in the English Subtask 2. **UZH** (Tang, 2020) submitted (after the competition had ended) a system based on the original BERT sense embeddings created for (Loureiro and Jorge, 2019) but improved their performance by combining them with contextualised embeddings. Finally for this group **AlexU-AUX-BERT** (Mahmoud and Torki, 2020) created new sense embeddings for the competition target words. In order to do so they sourced additional contexts for the top WordNet synsets. Their system scored third in the English Subtask 2. The pretrained WordNet sense embedding proved highly successful in this task, especially in Subtask 2, predicting the similarity scores themselves. The biggest weakness of the approach is their reliance on linguistic resources that don't exist for most languages other than English.

Related to these systems, the submission by **MineriaUNAM** (Gomez-Adorno et al., 2020) won the English Subtask 2. They proposed a system in which they calculated K-Means inspired centroids from the words in the context and used them to modify the original SimLex-999 non contextualised similarity scores. The approach, even if very successful, seems to rely on having out of context human annotations, perhaps not realistic in the general case. The fact that the system did very poorly in Subtask 1, which asked to predict change, seems to indicate much of the success is coming from the human annotations. A related strategy could perhaps be used with embeddings or computed predictions instead of human scores.

The next group focused on testing a variety of models and parameters. **BRUMS** (Hettiarachchi and Ranasinghe, 2020) worked with ELMo, BERT, Flair (Akbik et al., 2018), Transformer-XL (Dai et al., 2019) and XLNet (Yang et al., 2019). Their final submission made use of stacked embeddings proposed by Akbik et al. (2018). They won the Finnish Subtask 2, ended second in the two Slovene ones and performed very well in the two English ones. The **Hitachi** team (Morishita et al., 2020) looked at BERT and XML-R. Their main insight was that for every language, the layers from the center to the end where always the best performing ones, however while BERT performed best in the last layer, XLM-R did in the center one, suggesting their inner structure is organised differently. They won the Slovene Subtask 1, finished second in the two Croatian subtasks and performed competitively in the English ones. To conclude with this group **JUSTMasters** (Al-Khdour et al., 2020) tested several models, parameters and their own strategy to combine models. They achieved very good performance, especially in the English Subtask 2. However, in order to optimise their system, they made many more submissions than allowed in the competition; we therefore leave them out of the official ranking.

With a more multilingual approach, **BabelEncoding** (Costella Pessutto et al., 2020) proposed a solution in which they translated the contexts and target words to many languages and then used a weighted combination of monolingual pretrained non contextualised embeddings and BERT embeddings. Their idea is that the translation not only brings new resources but the process itself can produce useful information, for example to disambiguate. The approach works very well for the less resourced languages, being clearly the best system in that category, in both Subtask 1 and 2. Their system won Subtask 1 and 2 for Croatian (by a healthy margin) and 2 for Slovene, ending third in the Slovene Subtask 1 and third and second in the two Finnish ones.

The **MultiSem** team (Soler and Apidianaki, 2020) collected 5 different datasets in order to fine-tune their BERT models, most of them automatically generated from previous datasets to increase contextual influence. As an example, ukWaC-subs was created by substituting target words by either: a correct substitute; a word that could be the right substitute in other circumstances but it is not in this context; or a random word. The datasets included WiC, which when used to fine tune the model resulted in the best performance for Subtask1, giving them a third place. The approach works very well, giving a very consistent performance in all categories, and significantly improving the non fine-tuned model from a ρ=0.715 and 0.661 per subtask, to a ρ=0.760 and 0.718 respectively.

Ferryman's focus (Chen et al., 2020) was clearly the English Subtask 1, which they won with a modification of BERT in which they fed the TF-IDF score of the words to the model, thus incorporating information about the general importance of words. The system does very well at predicting the change between contexts, but surprisingly poorly at predicting similarity itself, ending last in the English Subtask 2 and second from the last in Croatian and Slovene.

The starting point of **CitiusNLP** (Gamallo, 2020) was the idea that, even if BERT seems to be able to encode syntactic structure, it doesn't seem to make use of it. They created a linguistically motivated system that relied in dependency to create predictions. However, its performance was considerably worse than BERT's and their actual submissions are based on a standard BERT model.

Finally, the **Will_Go** team (Bao et al., 2020) looked at different ways to measure similarity between embeddings, mixing euclidean distance with the most common cosine similarity and several others not described in their paper. The combination works well, they achieved a second place in the English Subtask 1 and won the Finnish Subtask 1.

8 Conclusion

We resented the SemEval-2020 Task on *Graded Word Similarity in Context* and introduced our new dataset *CoSimLex*. We provided the motivation behind their design choices and described the annotation process. The task received a good number of submissions and system description papers (15 and 11 respectively). We hope both the task and the dataset will be useful for researchers looking into how state-of-the-art systems capture context, and help promote the use of psychologically and cognitively inspired ideas in our field. Some of the interesting highlights were good performance of WordNet-based sense embeddings, the improvements achieved in less-resourced languages by simply translating the input, how the explicit feeding of an "old-fashioned" feature like TF-IDF improved a very modern system's performance, and the power of well designed, automatically created datasets for fine-tuning.

Additional and more detailed analyses of the dataset and task results will follow as part of future work. Areas to be investigated include the impact of different similarity ranges and degrees of polysemy, and more detailed qualitative analysis of the differences in annotation and between systems.

Acknowledgements

This research is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains. Carlos S. Armendariz is also supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1); Matthew Purver is also supported by the EPSRC project Streamlining Social Decision Making for Improved Internet Standards (SoDe- Stream, EP/S033564/1). The work of Ivan Vulić is supported by the ERC Consolidator Grant LEXICAL (no. 648909).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Nour Al-Khdour, Mutaz Bni Younes, Malak Abdullah, and AL-Smadi Mohammad. 2020. JUSTMasters at SemEval-2020 Task 3: Multilingual deep learning model to predict the effect of context in word similarity. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Carlos S. Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France, May. European Language Resources Association.

Wei Bao, Hongshu Che, and Jiandong Zhang. 2020. Will_Go at SemEval-2020 Task 3: An accurate model for predicting the (graded) effect of context in word similarity based on BERT. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Weilong Chen, Xin Yuan, Sai Zhang, Jiehui Wu, Yanru Zhang, and Yang Wang. 2020. Ferryman at SemEval-2020 Task: BERT with TFIDF-weighting for predicting the effect of context in word similarity. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, F. Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

Lucas Rafael Costella Pessutto, Viviane P. Moreira, Tiago de Melo, and Altigran da Silva. 2020. BabelEncoding at SemEval-2020 Task 3: Contextual similarity as a combination of multilingualism and language models. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

David A. Cruse. 1986. *Lexical semantics*. Cambridge university press.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Vyvyan Evans and Melanie Green. 2018. *Cognitive Linguistics: An Introduction*. Routledge.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Pablo Gamallo. 2020. CitiusNLP at SemEval-2020 Task 3: Comparing two approaches for word vector contextualization. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Peter Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 Shared Task - automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Helena Gomez-Adorno, Gemma Bel-Enguix, Jorge Reyes-Magaña, Benjamin Moreno, Ramon Casillas, and Daniel Vargas. 2020. MineriaUNAM at SemEval-2020 Task 3: Predicting contextual word similarity using a centroid based approach and word embeddings. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2020. BRUMS at SemEval-2020 Task 3: Contextualised embeddings for predicting the (graded) effect of context in word similarity. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Claudia Kittask. 2019. *Computational Models of Concept Similarity for the Estonian Language*. Bachelor's thesis, University of Tartu.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy, July. Association for Computational Linguistics.

Somaia Mahmoud and Marwan Torki. 2020. AlexU-AUX-BERT at SemEval-2020 Task 3: Improving BERT contextual similarity using multiple auxiliary contexts. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

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

Terufumi Morishita, Gaku Morio, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 Task 3: Exploring the representation spaces of transformers for human sense word similarity. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

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.

Aina Gari Soler and Marianna Apidianaki. 2020. MultiSem at SemEval-2020 Task 3: Fine-tuning BERT for lexical meaning. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Li Tang. 2020. UZH at SemEval-2020 Task 3: Combining BERT with WordNet sense embeddings to predict graded word similarity changes. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Matej Ulčar and Marko Robnik-Šikonja. 2020. High quality ELMo embeddings for seven less-resourced languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4731–4738, Marseille, France, May. European Language Resources Association.

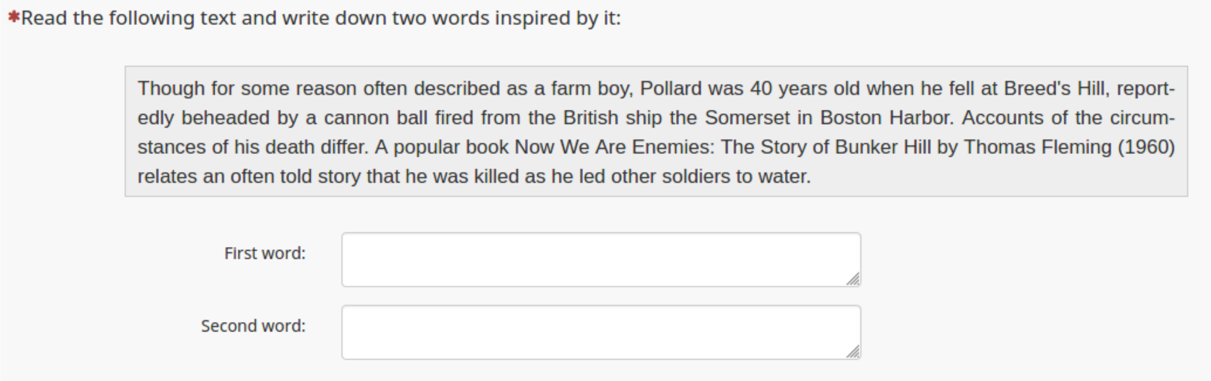
Viljami Venekoski and Jouko Vankka. 2017. Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131 in Linköping Electronic Conference Proceedings, pages 231–236. Linköping University Electronic Press, Linköpings universitet.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *arXiv preprint arXiv:2003.04866*.

Massimo Warglien and Peter Gärdenfors. 2015. Meaning negotiation. In *Applications of conceptual spaces*, pages 79–94. Springer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

A Appendix: Survey Example



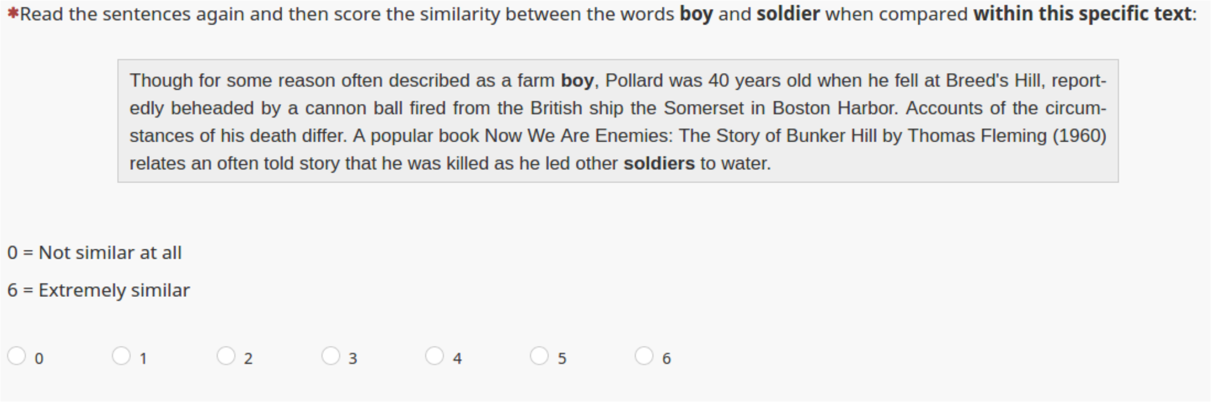
*Read the following text and write down two words inspired by it:

Though for some reason often described as a farm boy, Pollard was 40 years old when he fell at Breed's Hill, reportedly beheaded by a cannon ball fired from the British ship the Somerset in Boston Harbor. Accounts of the circumstances of his death differ. A popular book Now We Are Enemies: The Story of Bunker Hill by Thomas Fleming (1960) relates an often told story that he was killed as he led other soldiers to water.

First word:

Second word:

Figure 3: First page shown for each word pair annotation task: annotators must read the context and come up with two words inspired by it. At this point, the word pair to be scored is not known to the annotator.



*Read the sentences again and then score the similarity between the words **boy** and **soldier** when compared **within this specific text**:

Though for some reason often described as a farm **boy**, Pollard was 40 years old when he fell at Breed's Hill, reportedly beheaded by a cannon ball fired from the British ship the Somerset in Boston Harbor. Accounts of the circumstances of his death differ. A popular book Now We Are Enemies: The Story of Bunker Hill by Thomas Fleming (1960) relates an often told story that he was killed as he led other **soldiers** to water.

0 = Not similar at all

6 = Extremely similar

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

Figure 4: Second page shown for each word pair annotation task: the same context is now shown with the target words in bold, and annotators must give a similarity score for the word pair within that particular context.

B Appendix: Less-resourced Examples

B.1 Croatian

Word1: nov **Word2: svjež**	**SimLex (English)**: μ 6.83 σ 1.2
Context1	**Context1:** μ 9.49 σ 1.05
U jesen 1175. Fridrik je zamolio **svježe** trupe iz Njemačke. Prije svega Henrik Lav kao najmoćniji knez i vladar Bavarske odbio je caru poslati **nove** vojnike uvjetujući to prepuštanjem Goslara s bogatim rudnicima srebra.	
Context2	**Context2:** μ 1.85 σ 2.42
Proučavanje upalnih promjena dokazao je da ulaženje bijelih krvnih tjelešaca u tkivo uzrokuje gnojenje. Po njegovoj teoriji, rak nastaje iz emrionalnih stanica, razbacanih po organizmu. Uveo je **nove** metode istraživanja, npr. smrzavanje **svježeg** tkiva i pravljenje mirkoskopskih rezova.	
	P-Value: 2.4×10^{-5}

Figure 5: Example from the Croatian dataset, showing a word pair with two contexts, each with mean and standard deviation of human similarity judgements. The P-Value shown is the result of a Mann-Whitney U test.

B.2 Slovene

Word1: zgodba **Word2: tema** **SimLex (English)**: μ 5 σ 1.7

Context1 **Context1:** μ 0.167 σ 0.527

V **zgodbi** Čajanka za psa mačka in papagaja, se cunjasta dvojčica Nina sooča s strahom. Ker je še majhna deklica se boji **teme**, toda na pomoč ji prihiti punčka in škratje Copatki, ki Nini predlagajo naj se poveselijo in priredijo čajanko. Skupaj s papagajem, psom in mačkom priredijo čajanko in pozabijo na strah.

Context2 **Context2:** μ 6.3 σ 1.11

Koreografijo je sestavil Jamal Sims, ki je z Miley Cyrus sodeloval že pri plesu za pesem »Hoedown Throwdown«. Miley Cyrus in Jamal Sims sta skupaj sestavila koreografijo, ki bi se ujemala z **zgodbo** v pesmi, in nazadnje vse skupaj predstavila Robertu Halsu, ki si je »takoj zamislil, kako bo vse skupaj izgledalo«. V zvezi s **temo** videospota je Miley Cyrus povedala: »Mislim, da videospot razloži, da moje življenje ne izključuje življenj drugih ljudi.

P-Value: 5.1×10^{-5}

Figure 6: Example from the Slovene dataset, showing a word pair with two contexts, each with mean and standard deviation of human similarity judgements. The P-Value shown is the result of a Mann-Whitney U test.

B.3 Finnish

Word1: rikos **Word2: varkaus** **SimLex (English)**: μ 7.53 σ 1.32

Context1 **Context1:** μ 4.33 σ 2.38

Valistuksen vaikutuksesta häpeärangaistuksista vähitellen luovuttiin. Esimodernissa Euroopassa häpeärangaistuksiin johtivat etupäässä pienehköt **rikokset**, kuten solvaukset ja häiritsevä juopumus, mutta myös esimerkiksi aviorikos ja **varkaus**. Häpeärangaistuksien toteuttamistavat vaihtelivat alueellisesti.

Context2 **Context2:** μ 0 σ 0

Tekoja voidaan siis pitää pääosin laittomina, koska tuolloin ei ollut käytettävissä kuolemanrangaistuksen sallivaa, asianmukaista lainsäädäntöä. Sisällissodan jälkeen laaditulla armahduslailla vapautettiin myös valkoisen osapuolen edustajat vastuusta mahdollisesti tekemistään **rikoksista**, joten jonkinlainen ymmärrys teloitusten laittomuudesta oli ollut olemassa jo tuolloin. Kuolemantuomioiden langettamista jatkoi **Varkauden** kenttäoikeus, jonka lainmukaisuudesta voidaan olla myös hyvin erimielisiä.

P-Value: 3.3×10^{-5}

Figure 7: Example from the Finnish dataset, showing a word pair with two contexts, each with mean and standard deviation of human similarity judgements. The P-Value shown is the result of a Mann-Whitney U test. This is a very particular example, while "rikos" translates as "crime" and "varkaus" as "theft", there is a town named "Varkaus", which is the meaning of the word in the second context. This is the reason why all the annotators, accurately scored the similarity of the two words as 0 in the second context.