Polish evaluation dataset for compositional distributional semantics models

Alina Wróblewska **Katarzyna Krasnowska-Kieraś**
Institute of Computer Science, Polish Academy of Sciences
alina@ipipan.waw.pl kasia.krasnowska@gmail.com

Abstract

The paper presents a procedure of building an evaluation dataset[1]. for the validation of compositional distributional semantics models estimated for languages other than English. The procedure generally builds on steps designed to assemble the SICK corpus, which contains pairs of English sentences annotated for semantic relatedness and entailment, because we aim at building a comparable dataset. However, the implementation of particular building steps significantly differs from the original SICK design assumptions, which is caused by both lack of necessary extraneous resources for an investigated language and the need for language-specific transformation rules. The designed procedure is verified on Polish, a fusional language with a relatively free word order, and contributes to building a Polish evaluation dataset. The resource consists of 10K sentence pairs which are human-annotated for semantic relatedness and entailment. The dataset may be used for the evaluation of compositional distributional semantics models of Polish.

1 Introduction and related work

1.1 Distributional semantics

The basic idea of distributional semantics, i.e. determining the meaning of a word based on its co-occurrence with other words, is derived from the empiricists – Harris (1954) and Firth (1957). John R. Firth drew attention to the context-dependent nature of meaning especially with his famous maxim "You shall know a word by the company it keeps" (Firth, 1957, p. 11).

Nowadays, distributional semantics models are estimated with various methods, e.g. word embedding techniques (Bengio et al., 2003, 2006; Mikolov et al., 2013). To ascertain the purport of a word, e.g. *bath*, you can use the context of other words that surround it. If we assume that the meaning of this word expressed by its lexical context is associated with a distributional vector, the distance between distributional vectors of two semantically similar words, e.g *bath* and *shower*, should be smaller than between vectors representing semantically distinct words, e.g. *bath* and *tree*.

1.2 Compositional distributional semantics

Based on empirical observations that distributional vectors encode certain aspects of word meaning, it is expected that similar aspects of the meaning of phrases and sentences can also be represented with vectors obtained via composition of distributional word vectors. The idea of semantic composition is not new. It is well known as the *principle of compositionality*:[2] "The meaning of a compound expression is a function of the meaning of its parts and of the way they are syntactically combined." (Janssen, 2012, p. 19).

Modelling the meaning of textual units larger than words using compositional and distributional information is the main subject of compositional distributional semantics (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Socher et al., 2012, to name a few studies). The fundamental principles of compositional distributional semantics, henceforth referred to as CDS, are mainly propagated with papers written on the topic. Apart from the papers, it was the SemEval-2014 Shared Task 1

[1] The dataset is obtainable at: http://zil.ipipan.waw.pl/Scwad/CDSCorpus

[2] As the principle of compositionality is attributed to Gottlob Frege, it is often called *Frege's principle*.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 784–792
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1073

(Marelli et al., 2014) that essentially contributed to the expansion of CDS and increased an interest in this domain. The goal of the task was to evaluate CDS models of English in terms of semantic relatedness and entailment on proper sentences from the SICK corpus.

1.3 The SICK corpus

The SICK corpus (Bentivogli et al., 2014) consists of 10K pairs of English sentences containing multiple lexical, syntactic, and semantic phenomena. It builds on two external data sources – the 8K ImageFlickr dataset (Rashtchian et al., 2010) and SemEval-2012 Semantic Textual Similarity dataset (Agirre et al., 2012). Each sentence pair is human-annotated for relatedness in meaning and entailment.

The relatedness score corresponds to the degree of semantic relatedness between two sentences and is calculated as the average of ten human ratings collected for this sentence pair on the 5-point Likert scale. This score indicates the extent to which the meanings of two sentences are related.

The entailment relation between two sentences, in turn, is labelled with *entailment*, *contradiction*, or *neutral*. According to the SICK guidelines, the label assigned by the majority of human annotators is selected as the valid entailment label.

1.4 Motivation and organisation of the paper

Studying approaches to various natural language processing (henceforth NLP) problems, we have observed that the availability of language resources (e.g. training or testing data) stimulates the development of NLP tools and the estimation of NLP models. English is undoubtedly the most prominent in this regard and English resources are the most numerous. Therefore, NLP methods are mostly designed for English and tested on English data, even if there is no guarantee that they are universal. In order to verify whether an NLP algorithm is adequate, it is not enough to evaluate it solely for English. It is also valuable to have high-quality resources for languages typologically different to English. Hence, we aim at building datasets for the evaluation of CDS models in languages other than English, which are often underresourced. We strongly believe that the availability of test data will encourage development of CDS models in these languages and allow to better test the universality of CDS methods.

We start with a high-quality dataset for Polish, which is a completely different language than English in at least two dimensions. First, it is a rather under-resourced language in contrast to the resource-rich English. Second, it is a fusional language with a relatively free word order in contrast to the isolated English with a relatively fixed word order. If some heuristics is tested on e.g. Polish, the evaluation results can be approximately generalised to other Slavic languages. We hope the Slavic NLP community will be interested in designing and evaluating methods of semantic modelling for Slavic languages.

The procedure of building an evaluation dataset for validating compositional distributional semantics models of Polish generally builds on steps designed to assemble the SICK corpus (described in Section 1.3) because we aim at building an evaluation dataset which is comparable to the SICK corpus. However, the implementation of particular building steps significantly differs from the original SICK design assumptions, which is caused by both lack of necessary extraneous resources for Polish (see Section 2.1) and the need for Polish-specific transformation rules (see Section 2.2). Furthermore, the rules of arranging sentences into pairs (see Section 2.3) are defined anew taking into account the characteristic of data and bidirectional entailment annotations, since an entailment relation between two sentences must not be symmetric. Even if our assumptions of annotating sentence pairs coincide with the SICK principles to a certain extent (see Section 3.1), the annotation process differs from the SICK procedure, in particular by introducing an element of human verification of correctness of automatically transformed sentences (see Section 3.2) and some additional post-corrections (see Section 3.3). Finally, a summary of the dataset is provided in Section 4.1 and the dataset evaluation is given in Section 4.2.

2 Procedure of collecting data

2.1 Selection and description of images

The first step of building the SICK corpus consisted in the random selection of English sentence pairs from existing datasets (Rashtchian et al., 2010; Agirre et al., 2012). Since we are not aware of accessibility of analogous resources for Polish, we have to select images first and then describe the selected images.

Images are selected from the 8K ImageFlickr

dataset (Rashtchian et al., 2010). At first we wanted to take only these images the descriptions of which were selected for the SICK corpus. However, a cursory check shows that these images are quite homogeneous, with a predominant number of dogs depictions. Therefore, we independently extract 1K images and split them into 46 thematic groups (e.g. *children*, *musical instruments*, *motorbikes*, *football*, *dogs*). The numbers of images within individual thematic groups vary from 6 images in the *volleyball* and *telephoning* groups to 94 images in the *various people* group. The second largest groups are *children* and *dogs* with 50 images each.

The chosen images are given to two authors who independently of each other formulate their descriptions based on a short instruction. The authors are instructed to write one single sentence (with a sentence predicate) describing the action in a displayed image. They should not describe an imaginable context or an interpretation of what may lie behind the scene in the picture. If some details in the picture are not obvious, they should not be described either. Furthermore, the authors should avoid multiword expressions, such as idioms, metaphors, and named entities, because those are not compositional linguistic phenomena. Finally, descriptions should contain Polish diacritics and proper punctuation.

2.2 Transformation of descriptions

The second step of building the SICK corpus consisted in pre-processing extracted sentences, i.e. normalisation and expansion (Bentivogli et al., 2014, p. 3–4). Since the authors of Polish descriptions are asked to follow the guidelines (presented in Section 2.1), the normalisation step is not essential for our data. The expansion step, in turn, is implemented and the sentences provided by the authors are lexically and syntactically transformed in order to obtain derivative sentences with similar, contrastive, or neutral meanings. The following transformations are implemented:

1. *dropping conjunction* concerns sentences with coordinated predicates sharing a subject, e.g. *Rowerzysta odpoczywa i obserwuje morze.* (Eng. ‘A cyclist is resting and watching the sea.’). The finite form of one of the coordinated predicates is transformed into:
    - an active adjectival participle, e.g. *Odpoczywający rowerzysta obserwuje morze.* (Eng. ‘A resting cyclist is watching the sea.’) or *Obserwujący morze rowerzysta odpoczywa.* (Eng. ‘A cyclist, who is watching the sea, is resting.’),
    - a contemporary adverbial participle, e.g. *Rowerzysta, odpoczywając, obserwuje morze.* (Eng. ‘A cyclist is watching the sea, while resting.’) or *Rowerzysta odpoczywa, obserwując morze.* (Eng. ‘A cyclist is resting, while watching the sea.’).
2. *removing conjunct in adjuncts*, i.e. the deletion of one of coordinated elements of an adjunct, e.g. *Mały, ale zwinny kot miauczy.* (Eng. ‘A small but agile cat miaows.’) can be changed into either *Mały kot miauczy.* (Eng. ‘A small cat miaows.’) or *Zwinny kot miauczy.* (Eng. ‘An agile cat miaows.’).
3. *passivisation*, e.g. *Człowiek ujeżdża byka.* (Eng. ‘A man is breaking a bull in.’) can be transformed into *Byk jest ujeżdżany przez człowieka.* (Eng. ‘A bull is being broken in by a man.’).
4. *removing adjuncts*, e.g. *Dwa białe króliki siedzą na trawie.* (Eng. ‘Two small rabbits are sitting on the grass.’) can be changed into *Króliki siedzą.* (Eng. ‘The rabbits are sitting.’).
5. *swapping relative clause for participles*, i.e. a relative clause swaps with a participle (and vice versa), e.g. *Kobieta przytula psa, którego trzyma na smyczy.* (Eng. ‘A woman hugs a dog which she keeps on a leash.’). The relative clause is interchanged for a participle construction, e.g. *Kobieta przytula trzymanego na smyczy psa.* (Eng. ‘A woman hugs a dog kept on a leash.’).
6. *negation*, e.g. *Mężczyźni w turbanach na głowach siedzą na słoniach.* (Eng. ‘Men in turbans on their heads are sitting on elephants.’) can be transformed into *Nikt nie siedzi na słoniach.* (Eng. ‘Nobody is sitting on elephants.’), *Żadni mężczyźni w turbanach na głowach nie siedzą na słoniach.* (Eng. ‘No men in turbans on their heads are sitting on elephants.’), and *Mężczyźni w turbanach na głowach nie siedzą na słoniach.* (Eng. ‘Men in turbans on their heads are not sitting on elephants.’).

7. *constrained mixing of dependents from various sentences*, e.g. *Dwoje dzieci siedzi na wielbłądach w pobliżu wysokich gór.* (Eng. 'Two children are sitting on camels near high mountains.') can be changed into *Dwoje dzieci siedzi przy zastawionym stole w pobliżu wysokich gór.* (Eng. 'Two children are sitting at the table laid with food near high mountains.').

The first five transformations are designed to produce sentences with a similar meaning, the sixth transformation outputs sentences with a contradictory meaning, and the seventh transformation should generate sentences with a neutral (or unrelated) meaning. All transformations are performed on the dependency structures of input sentences (Wróblewska, 2014).

Some of the transformations are very productive (e.g. mixing dependents). Other, in turn, are sparsely represented in the output (e.g. dropping conjunction). The number of transformed sentences randomly selected to build the dataset is in the second column of Table 1.

transformation	**selected**	
dropping conjunction	139	2.0%
removing conjunct in adjunct	485	6.9%
passivisation	893	12.8%
removing adjuncts	1013	14.5%
swapping rc↔ptcp	1291	18.4%
negation	1304	18.6%
mixing dependents	1878	26.8%

Table 1: Numbers of transformed sentences selected for annotation.

2.3 Data ensemble

The final step of building the SICK corpus consisted in arranging normalised and expanded sentences into pairs. Since our data diverges from SICK data, the process of arranging Polish sentences into pairs also differs from pairing in the SICK corpus. The general idea behind the pair-ensembling procedure was to introduce sentence pairs with different levels of relatedness into the dataset. Apart from pairs connecting two sentences originally written by humans (as described in Section 2.1), there are also pairs in which an original sentence is connected with a transformed sentence. For each of the 1K images, the following 10 pairs are constructed (for A being the set of all sentences originally written by the first author, B being the set of all sentences originally written by the second author, $\mathbf{a} \in A$ and $\mathbf{b} \in B$ being the original descriptions of the picture):

1. $(\mathbf{a}, \mathbf{b})$
2. $(\mathbf{a}, \mathbf{a}_1)$, where $\mathbf{a}_1 \in t(\mathbf{a})$, and $t(\mathbf{a})$ is the set of all transformations of the sentence $\mathbf{a}$
3. $(\mathbf{b}, \mathbf{b}_1)$, where $\mathbf{b}_1 \in t(\mathbf{b})$
4. $(\mathbf{a}, \mathbf{b}_2)$, where $\mathbf{b}_2 \in t(\mathbf{b})$
5. $(\mathbf{b}, \mathbf{a}_2)$, where $\mathbf{a}_2 \in t(\mathbf{a})$
6. $(\mathbf{a}, \mathbf{a}_3)$, where $\mathbf{a}_3 \in t(\mathbf{a}'), \mathbf{a}' \in A, \mathcal{T}(\mathbf{a}') = \mathcal{T}(\mathbf{a}), \mathbf{a}' \neq \mathbf{a}$, for $\mathcal{T}(\mathbf{a})$ being the thematic group[3] of $\mathbf{a}$
7. $(\mathbf{b}, \mathbf{b}_3)$, where $\mathbf{b}_3 \in t(\mathbf{b}'), \mathbf{b}' \in B, \mathcal{T}(\mathbf{b}') = \mathcal{T}(\mathbf{b}), \mathbf{b}' \neq \mathbf{b}$
8. $(\mathbf{a}, \mathbf{a}_4)$, where $\mathbf{a}_4 \in A, \mathcal{T}(\mathbf{a_4}) \neq \mathcal{T}(\mathbf{a})$[4]
9. $(\mathbf{b}, \mathbf{b}_4)$, where $\mathbf{b_4} \in B, \mathcal{T}(\mathbf{b_4}) \neq \mathcal{T}(\mathbf{b})$
10. $(\mathbf{a}, \mathbf{a}_5)$, where $\mathbf{a}_5 \in t(\mathbf{a}), \mathbf{a}_5 \neq \mathbf{a}_1$ for 50% images, $(\mathbf{b}, \mathbf{b}_5)$ (analogously) for other 50%.[5]

For each sentence pair $(\mathbf{a}, \mathbf{b})$ created according to this procedure, its reverse $(\mathbf{b}, \mathbf{a})$ is also included in our corpus. As a result, the working set consists of 20K sentence pairs.

3 Corpus annotation

3.1 Annotation assumptions

The degree of semantic relatedness between two sentences is calculated as the average of all human ratings on the Likert scale with the range from 0 to 5. Since we do not want to excessively influence

[3]The thematic group of a sentence $\mathbf{a}$ corresponds to the thematic group of an image being the source of $\mathbf{a}$ (as described in Section 2.1).

[4]The pairs $(\mathbf{a}, \mathbf{a}_4)$ of the same authors' descriptions of two images from different thematic groups are expected to be unrelated. The same applies to $(\mathbf{b}, \mathbf{b}_4)$.

[5]A repetition of point 2 with a restriction that a different pair is created (pairs of very related sentences are expected). We alternate between authors A and B to obtain equal author proportions in the final ensemble of pairs.

the annotations, the guidelines given to annotators are mainly example-based:[6]

- 5 (very related): *Kot siedzi na płocie.* (Eng. 'A cat is sitting on the fence.') vs. *Na płocie jest duży kot.* (Eng. 'There is a large cat on the fence.'),

- 1–4 (more or less related):
  Kot siedzi na płocie. (Eng. 'A cat is sitting on the fence.') vs. *Kot nie siedzi na płocie.* (Eng. 'A cat is not sitting on the fence.');
  Kot siedzi na płocie. (Eng. 'A cat is sitting on the fence.') vs. *Właściciel dał kotu chrupki.* (Eng. 'The owner gave kibble to his cat.');
  Kot siedzi na płocie. (Eng. 'A cat is sitting on the fence.') vs. *Kot miauczy pod płotem.* (Eng. 'A cat miaows by the fence.').

- 0 (unrelated): *Kot siedzi na płocie.* (Eng. 'A cat is sitting on the fence.') vs. *Zaczął padać deszcz.* (Eng. 'It started to rain.').

Apart from these examples, there is a note in the annotation guidelines indicating that the degree of semantic relatedness is not equivalent to the degree of semantic similarity. Semantic similarity is only a special case of semantic relatedness, semantic relatedness is thus a more general term than the other one.

Polish entailment labels correspond directly to the SICK labels (i.e. *entailment*, *contradiction*, *neutral*). The entailment label assigned by the majority of human judges is selected as the gold label. The entailment labels are defined as follows:

- **a wynika z b** (**b** entails **a**) – if a situation or an event described by sentence **b** occurs, it is recognised that a situation or an event described by **a** occurs as well, i.e. **a** and **b** refer to the same event or the same situation,

- **a jest zaprzeczeniem b** (**a** is the negation of **b**) – if a situation or an event described by **b** occurs, it is recognised that a situation or an event described by **a** may not occur at the same time,

- **a jest neutralne wobec b** (**a** is neutral to **b**) – the truth of a situation described by **a** cannot be determined on the basis of **b**.

3.2 Annotation procedure

Similar to the SICK corpus, each Polish sentence pair is human-annotated for semantic relatedness and entailment by 3 human judges experienced in Polish linguistics.[7] Since for each annotated pair $(\mathbf{a}, \mathbf{b})$, its reverse $(\mathbf{b}, \mathbf{a})$ is also subject to annotation, the entailment relation is in practice determined 'in both directions' for 10K sentence pairs. For the task of relatedness annotation, the order of sentences within pairs seems to be irrelevant, we can thus assume to obtain 6 relatedness scores for 10K unique pairs.

Since the transformation process is fully automatic and to a certain extent based on imperfect dependency parsing, we cannot ignore errors in the transformed sentences. In order to avoid annotating erroneous sentences, the annotation process is divided into two stages:

1. a sentence pair is sent to a judge with the *leader* role, who is expected to edit and to correct the transformed sentence from this pair before annotation, if necessary,

2. the verified and possibly enhanced sentence pair is sent to the other two judges, who can only annotate it.

The *leader* judges should correct incomprehensible and ungrammatical sentences with a minimal number of necessary changes. Unusual sentences which could be accepted by Polish speakers should not be modified. Moreover, the modified sentence may not be identical with the other sentence in the pair. The classification and statistics of distinct corrections made by the *leader* judges are provided in Table 2.

A strict classification of error types is quite hard to provide because some sentences contain more than one error. We thus order the error types from the most serious errors (i.e. 'sense' errors) to the redundant corrections (i.e. 'other' type). If a sentence contains several errors, it is qualified for the higher order error type.

In the case of sentences with 'sense' errors, the need for correction is uncontroversial and

[6] We realise that the boundary between semantic perception of a sentence by various speakers is fuzzy (it depends on speakers' education, origin, age, etc.). It was thus our well-thought-out decision to draw only general annotation frames and to enable annotators to rely on their feel for language.

[7] Our annotators have relatively strong linguistic background. Five of them have PhD in linguistics, five are PhD students, one is a graduate, and one is an undergraduate.

error type	# of errors	% of errors
sense	171	12.3
semantic	407	29.2
grammatical	243	17.4
word order	141	10.1
punctuation	366	26.2
other	68	4.9

Table 2: Classification and statistics of corrections.

arises from an internal logical contradiction.[8] The sentences with 'semantic' changes are syntactically correct, but deemed unacceptable by the *leader* annotators from the semantic or pragmatic point of view.[9] The 'grammatical' errors mostly concern missing agreement.[10] The majority of 'word order' corrections are unnecessary, but we found some examples which can be classified as actual word or phrase order errors.[11] The correction of punctuation consists in adding or deleting a comma.[12] The sentences in the 'other' group, in turn, could as well have been left unchanged because they are proper Polish sentences, but were apparently considered odd by the *leader* annotators.

[8] An example of 'sense' error: the sentence *Chłopak w zielonej bluzie i czapce zjeżdża na rolkach* ***na leżąco***. (Eng. 'A boy in a green sweatshirt and a cap roller-skates downhill **in a lying position**.') is corrected into *Chłopak w zielonej bluzie i czapce zjeżdża na rolkach.* (Eng. 'A boy in a green sweatshirt and a cap roller-skates downhill.').

[9] An example of 'semantic' correction: the sentence *Dziewczyna trzyma w* ***pysku*** *patyk.* (Eng. 'A girl holds a stick in her **muzzle**.') is corrected into *Dziewczyna trzyma w* ***ustach*** *patyk.* (Eng. 'A girl holds a stick in her **mouth**.').

[10] An example of 'grammatical' error: the sentence *Grupa*$_{\mathbf{sg}.nom}$ *uśmiechających się ludzi* ***tańczą***$_{pl}$. (Eng. *'A group of smiling people are dancing.') is corrected into *Grupa*$_{\mathbf{sg}.nom}$ *uśmiechających się ludzi* ***tańczy***$_{sg}$. (Eng. 'A group of smiling people is dancing.').

[11] An example of word order error: the sentence *Samochód, który jest uszkodzony,* ***koloru białego*** *stoi na lawecie dużego auta.* (lit. 'A car that is damaged, **of the white color** stands on the trailer of a large car.', Eng. 'A **white** car that is damaged is standing on the trailer of a large car.') is corrected into *Samochód* ***koloru białego****, który jest uszkodzony, stoi na lawecie dużego auta.*

[12] An example of punctuation correction: the wrong comma in the sentence *Nad brzegiem wody, stoją dwaj mężczyźni z wędkami.* (lit. 'On the water's edge, two men are standing with rods.'; Eng. 'Two men with rods are standing on the water's edge.') should be deleted, i.e. *Nad brzegiem wody stoją dwaj mężczyźni z wędkami.*

3.3 Impromptu post-corrections

During the annotation process it came out that sentences accepted by some human annotators are unacceptable for other annotators. We thus decided to garner annotators' comments and suggestions for improving sentences. After validation of these suggestions by an experienced linguist, it turns out that most of these proposals concern punctuation errors (e.g. missing comma) and typos in 312 distinct sentences. These errors are fixed directly in the corpus because they should not impact the annotations of sentence pairs. The other suggestions concern more significant changes in 29 distinct sentences (mostly minor grammatical or semantic problems overlooked by the *leader* annotators). The annotations of pairs with modified sentences are resent to the annotators so that they can verify and update them.

4 Corpus summary and evaluation

4.1 Corpus statistics

Tables 3 and 4 summarise the annotations of the resulting 10K sentence pairs corpus. Table 3 aggregates the occurrences of 6 possible relatedness scores, calculated as the mean of all 6 individual annotations, rounded to an integer.

relatedness	# of pairs
0	1978
1	1428
2	1082
3	2159
4	2387
5	966

Table 3: Final relatedness scores rounded to integers (total: 10K pairs).

Table 4 shows the number of the particular entailment labels in the corpus. Since each sentence pair is annotated for entailment in both directions, the final entailment label is actually a pair of two labels:

- *entailment+neutral* points to 'one-way' entailment,
- *contradiction+neutral* points to 'one-way' contradiction,
- *entailment+entailment*, *contradiction+contradiction*, and *neutral+neutral* point to equivalence.

While the actual corpus labels are ordered in the sense that there is a difference between e.g. *entailment+neutral* and *neutral+entailment* (the entailment occurs in different directions), we treat all labels as unordered for the purpose of this summary (e.g. *entailment+neutral* covers *neutral+entailment* as well, representing the same type of relation between two sentences).

entailment	**# of pairs**
neutral+neutral	6483
entailment+neutral	1748
entailment+entailment	933
contradiction+contradiction	721
contradiction+neutral	115

Table 4: Final entailment labels (total: 10K pairs).

4.2 Inter-annotator agreement

The standard measure of inter-annotator agreement in various natural language labelling tasks is Cohen's kappa (Cohen, 1960). However, this coefficient is designed to measure agreement between two annotators only. Since there are three annotators of each pair of ordered sentences, we decided to apply Fleiss' kappa[13] (Fleiss, 1971) designed for measuring agreement between multiple raters who give categorical ratings to a fixed number of items. An additional advantage of this measure is that different items can be rated by different human judges, which doesn't impact measurement. The normalised Fleiss' measure of inter-annotator agreement is:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where the quantity $\bar{P} - \bar{P}_e$ measures the degree of agreement actually attained in excess of chance, while "[t]he quantity $1 - \bar{P}_e$ measures the degree of agreement attainable over and above what would be predicted by chance" (Fleiss, 1971, p. 379).

We recognise Fleiss' kappa as particularly useful for measuring inter-annotator agreement with respect to entailment labelling in our evaluation dataset. First, there are more than two raters. Second, entailment labels are categorically. Measured with Fleiss' kappa, there is an inter-annotator agreement of $\kappa = 0.734$ for entailment labels in Polish evaluation dataset, which is quite satisfactory as for a semantic labelling task.

Relative to semantic relatedness, the distinction in meaning of two sentences made by human judges is often very subtle. This is also reflected in the inter-annotator agreement scores measured with Fleiss' kappa. Inter-annotator agreement measured for six semantic relatedness groups corresponding to points on the Likert scale is quite low: $\kappa = 0.337$. If we measure inter-annotator agreement for three classes corresponding to the three relatedness groups from the annotation guidelines (see Section 3.1), i.e. <0>, <1, 2, 3, 4>, and <5>, the Fleiss' score is significantly higher: $\kappa = 0.543$. Hence, we conclude that Fleiss' kappa is not a reliable measure of inter-annotator agreement in relation to relatedness scores. Therefore, we decided to use Krippendorff's α instead.

Krippendorff's α (Krippendorff, 1980, 2013) is a coefficient appropriate for measuring the inter-annotator agreement of a dataset which is annotated with multiple judges and characterised by different magnitudes of disagreement and missing values. Krippendorff proposes distance metrics suitable for various scales: binary, nominal, interval, ordinal, and ratio. In ordinal measurement[14] the attributes can be rank-ordered, but distances between them do not have any meaning. Measured with Krippendorff's ordinal α, there is an inter-annotator agreement of $\alpha = 0.780$ for relatedness scores in the Polish evaluation dataset, which is quite satisfactory as well. Hence, we conclude that our dataset is a reliable resource for the purpose of evaluating compositional distributional semantics model of Polish.

5 Conclusions

The goal of this paper is to present the procedure of building a Polish evaluation dataset for the validation of compositional distributional semantics models. As we aim at building an evalua-

[13] As Fleiss' kappa is actually the generalisation of Scott's π (Scott, 1955), it is sometimes referred to as Fleiss' *multi-*π, cf. Artstein and Poesio (2008).

[14] Nominal measurement is useless for measuring agreement between relatedness scores ($\alpha = 0.340$ is the identical value as Fleiss' kappa, since all disagreements are considered equal). We also test interval measurement, in which the distance between the attributes does have meaning and an average of an interval variable is computed. The interval score measured for relatedness annotations is quite high $\alpha = 0.785$, but we doubt whether the distance between relatedness scores is meaningful in this case.

tion dataset which is comparable to the SICK corpus, the general assumptions of our procedure correspond to the design principles of the SICK corpus. However, the procedure of building the SICK corpus cannot be adapted without modifications. First, the Polish seed-sentences have to be written based on the images which are selected from 8K ImageFlickr dataset and split into thematic groups, since usable datasets are not publicly available. Second, since the process of transforming sentences seems to be language-specific, the linguistic transformation rules appropriate for Polish have to be defined from scratch. Third, the process of arranging Polish sentences into pairs is defined anew taking into account the data characteristic and bidirectional entailment annotations. The discrepancies relative to the SICK procedure also concern the annotation process itself. Since an entailment relation between two sentences must not be symmetric, each sentence pair is annotated for entailment in both directions. Furthermore, we introduce an element of human verification of correctness of automatically transformed sentences and some additional post-corrections.

The presented procedure of building a dataset was tested on Polish. However, it is very likely that the annotation framework will work for other Slavic languages (e.g. Czech with an excellent dependency parser).

The presented procedure results in building the Polish test corpus of relatively high quality, confirmed by the inter-annotator agreement coefficients of $\kappa = 0.734$ (measured with Fleiss' kappa) for entailment labels and of $\alpha = 0.780$ (measured with Krippendorff's ordinal alpha) for relatedness scores.

Acknowledgments

We would like to thank the reliable and tenacious annotators of our dataset: Alicja Dziedzic-Rawska, Bożena Itoya, Magdalena Król, Anna Latusek, Justyna Małek, Małgorzata Michalik, Agnieszka Norwa, Małgorzata Szajbel-Keck, Alicja Walichnowska, Konrad Zieliński, and some other. The research presented in this paper was supported by SONATA 8 grant no 2014/15/D/HS2/03486 from the National Science Centre Poland.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*. pages 385–393.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34:557–596.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pages 1183–1193.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3:1137–1155.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural Probabilistic Language Models. In D.E. Holmes and L.C. Jain, editors, *Innovations in Machine Learning. Theory and Applications*, Springer-Verlag, Berlin Heidelberg, volume 194 of *Studies in Fuzziness and Soft Computing*, pages 137–186.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2014. SICK through the SemEval Glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Journal of Language Resources and Evaluation* 50:95–124.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

John Rupert Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis. Special volume of the Philological Society* pages 1–32.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 75:378–382.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. pages 1394–1404.

Zellig Harris. 1954. Distributional structure. *Word* 10:146–162.

Theo M. V. Janssen. 2012. Compositionality: its historic context. In Wolfram Hinzen, Edouard Machery, and Markus Werning, editors, *The Oxford Handbook of Compositionality*, Oxford University Press, Studies in Fuzziness and Soft Computing, pages 19–46.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage Publication, Thousand Oaks, 3rd edition.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pages 1–8.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26. Proceedings of Neural Information Processing Systems 2013*. pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34:1388–1429.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. pages 139–147.

William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 19:321–325.

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 1201–1211.

Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.