KaMiKla at SemEval-2022 Task 3: AlBERTo, BERT, and CamemBERT—Be(r)tween Taxonomy Detection and Prediction

Karl Vetter
karl.vetter@student.uni-tuebingen.de

Miriam Segiet
miriam.segiet@student.uni-tuebingen.de

Klara Lennermann
klara.lennermann@student.uni-tuebingen.de

Abstract

This paper describes our system submitted for SemEval Task 3: Presupposed Taxonomies: Evaluating Neural Network Semantics (Zamparelli et al., 2022). We participated in both the binary classification and the regression subtask. Target sentences are classified according to their taxonomical relation in subtask 1 and according to their acceptability judgment in subtask 2. Our approach in both subtasks is based on a neural network BERT model. We used separate models for the three languages covered by the task, English, French, and Italian. For the second subtask, we used median averaging to construct an ensemble model. We ranked 15th out of 21 groups for subtask 1 (F1-score: 77.38%) and 11th out of 17 groups for subtask 2 (RHO: 0.078).

1 Introduction

The recognition of lexical relationships between words and the corresponding generalization has attracted increasing attention in computational linguistics. Today, there already exist resources covering manually marked semantic relationships, e.g., taxonomic relations, such as the lexical database WordNet (Miller, 1992) or the multilingual dictionary and semantic network BabelNet (Navigli and Ponzetto, 2010).

Luu et al. (2016) define taxonomic relations between two terms as an is–a relation. In such a relation there is a hypernym, i.e., a supertype, and a hyponym, i.e., a subtype. Both the supertype and subtype are sets covering, in our task, certain semantic categories. In a relation such as *animal–dog*, the *animal* is the superordinate category and the *dog* is the subordinate term. Furthermore, those specific sets are included in a relation forming a special hierarchy (Kay, 1971). As stated by Nguyen et al. (2017), in such an is–a relation the supertype necessarily implies the subtype, but not vice versa.

SemEval 2022 Task 3 is a taxonomy detection and prediction task consisting of two subtasks: a binary classification and a regression task, both covering the languages English, French, and Italian.

We propose an approach based on the transformer-based machine learning model BERT. Since BERT is a bidirectional model producing state-of-the-art results (Devlin et al., 2019), we used this pre-trained model for our analysis. For the three different languages, the corresponding BERT models were used (Devlin et al., 2019; Polignano et al., 2019; Martin et al., 2020).[1]

2 Task Description

In the present shared task (Zamparelli et al., 2022), the taxonomic sentence structures in the given files are composed of different artificially generated constructions enforcing presuppositions.

Table 1 shows example sentences provided in the English test set. According to the task description page, the French and Italian datasets are translated versions of the English dataset that were slightly adapted.

The argument nouns in the given files come from 30 semantic categories including, among others, dogs, birds, and mammals. The given word sets already show broader and narrower categories (*mammals* vs. *dogs*/*birds*). Nevertheless, as shown in the examples in table 1, not all pairs of nouns reflect such a superordinate–subordinate relation (*apple* vs. *cauliflower*) as described by Nguyen et al. (2017). Therefore, the taxonomies do not only represent a direct is–a relation such that one given nominal is the subcategory and the other one is the supercategory. Thus, it has to be considered that human language consists of many argument and sentence structures that restrict such relations. That means the sentence structures also cover comparisons where both nouns come from

[1]The source code of our model is available at https://github.com/cicl-iscl/SemEval3.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 282 - 290

Construction	Example
andtoo	I like teaspoons, and mugs too.
butnot	I like cats, but not frogs
comparatives	I like apples as much as cauliflower.
drather	I would rather have veal than salmon.
except	I like seafood, with the only exception of salmon.
generally	I like peaches, and more generally fruits.
particular	I like jewelry, and in particular necklaces.
prefer	I do not like tiramisu, I prefer broccoli.
type	I like parrots, not other types of birds.
unlike	Unlike glass, PVC is often mentioned in this text.

Table 1: Example sentences from subtask 1 with a binary label 1 (i.e., acceptable).

the same broader category. This has also been covered in the work by Clarke (2012), who refers to taxonomies as a framework represented in a hierarchy where lexical counterparts or synonyms are considered. Therefore, the given shared task comes with a challenge different from only recognizing the taxonomic relation—furthermore, the embedding construction allowing or disallowing the given relation had to be checked.

The participants were provided with two datasets to work with the individual subtasks. The training set for subtask 1 was composed of 5,837 sentences for each language with binary labels representing 1 as an acceptable sentence and 0 as an unacceptable sentence. This subtask covers the binary prediciton of acceptability labels of each sentence given in the test set with 14,560 samples. Subtask 2 consists of the prediction of an average score on a seven-point Likert scale for 1,009 sentences in the test set. The original scores in the training set were annotated by humans.

Figure 1 shows the scores assigned to the sentences presented in the training set in subtask 2. This figure only shows the scores averaged over all annotators since per-annotator information is not available.

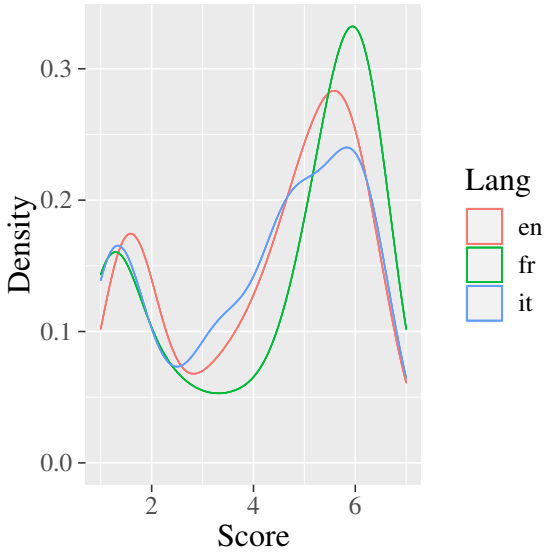

Figure 1: Distribution of the scores given to the sentences from the training set of subtask 2 by annotators.

3 System Overview

We used pre-trained BERT (Devlin et al., 2019) networks for subtasks 1 and 2. Separate models were used for each of the languages, specifically, the AlBERTo model (Polignano et al., 2019) was used for Italian, BERT base uncased was used for English, and the monolingual CamemBERT (Martin et al., 2020) model was used for French. The BERT models are powerful and highly versatile language models that possess the benefit of having learned good representations of the language they were trained on. This gives them a decisive edge over using models that are trained exclusively with the data provided for training as these pre-trained models will, for example, have encountered and learned representations for words that are not in the training set but are in the test set, whereas a model trained only on the training set will have trouble dealing with these unfamiliar words. As such they offer the opportunity for better generalization.

The bigger challenge of these tasks was not to produce models that perform well on the limited training data but to produce models that generalize well and do not merely overfit on the provided data. To this end, standard deep learning regularization techniques such as weight decay, dropout, and model averaging were used; nonetheless, the models performed much worse on the test data than on the validation data. Actually producing models that perform better at generalizing would likely

have required data augmentation and/or alternative training routines.

The models were not shared between subtasks 1 and 2, meaning that while the same pre-trained models were used for subtasks 1 and 2, the model used for subtask 2 was not fine-tuned for subtask 1 and vice versa.

3.1 Subtask 1

For subtask 1, the sentences were tokenized using the tokenizers of the pre-trained BERT models. The BERT model was extended with one fully connected hidden layer and an output layer. The model was trained to perform the classification task using cross-entropy loss, backpropagated using the Adamw optimizer (Loshchilov and Hutter, 2017), which combines the Adam optimizer (Kingma and Ba, 2017) with weight decay regularization. Dropout was used for further regularization. Gradient clipping by-norm was applied to solve the exploding gradient problem. Thirty percent of the data was used as a validation set, learning was terminated through early stopping with the loss function as the stopping criterion.

3.2 Subtask 2

The model for subtask 2 is similar to the model for task 1, again extending the BERT models with a hidden layer and one output layer producing a single number for the regression task. The output and targets were normalized to be between zero and one. The inverse transformation was then applied to the model output to get the final output on the original scale. The model was trained on seventy percent of the training data using mean squared error as the loss function. Dropout was used and weight decay was applied through the AdamW optimizer. The training was terminated using early stopping with the remaining thirty percent of the training data used for validation. For this task, we trained ten models per language, each with its own training split of the data. The final prediction for the test set was the median prediction of these models. We chose to use the median and not the mean as it is less affected by outlier predictions.

3.3 Hyperparameters

Hyperparameters were determined using grid search over a limited selection of plausible candidate values, including learning rate (1×10^{-5} for all models), the number of fully connected layers, neurons per layer, and in the case of subtask 2 a multiclass approach was also tested. For subtask 1 choosing bigger models, in the end, 2 hidden layers with 512 neurons each were used, which led to improvements on the validation set but may have caused overfitting that negatively impacted performance on the test set. For subtask 2, bigger models did not perform better than small ones, and as a result, the final models contained only a single hidden layer with 16 neurons. The complete hyperparameters are listed in appendices A and C.

	total	**It**	**Fr**	**En**
Precision	0.75	0.73	0.77	0.74
Recall	0.80	0.73	0.89	0.80
F1	0.78	0.73	0.83	0.77

Table 2: Results for subtask 1. All scores are micro-averaged over different constructions. The total scores are also micro-averaged over the languages.

4 Results

Unless stated otherwise, we analyzed, evaluated, and visualized the results in R (R Core Team, 2020) with the help of the packages caret (Kuhn, 2021), dplyr (Wickham et al., 2021), ggplot2 (Wickham, 2016), plyr (Wickham, 2011), readr (Wickham and Hester, 2020), stringr (Wickham, 2019), tikzDevice (Sharpsteen and Bracken, 2020), tm (Feinerer et al., 2008), and xtable (Dahl et al., 2019).

4.1 Subtask 1

The test data contains 14,560 sentences per language. Table 2 shows the results of the evaluation for each language. Out of the 18 participating teams (and three additional teams who only submitted results for the English subset), the KaMiKla models ranked 15th in the overall competition and the Italian part and 12th and 16th in the French and English portion, respectively.

The highest overall score in the competition was 0.94, an F1 measure of 0.93 for Italian and French, and 0.97 for English. The trivial baseline that used n-grams as features reached a global score of 0.73. The F1 score for the Italian model is 0.68, 0.76 for French, and 0.73 for English.

After the evaluation phase ended, the test sets were published containing an additional label that denoted the syntactic construction used in the respective sentence. Table 1 illustrates examples of what the labels mean. Further analysis revealed that the type of construction had an enormous effect on

the performance of the KaMiKla models. Figure 2 gives an overview of this. For the detailed evaluation metrics for each language by construction, see Appendix B.

The KaMiKla models performed with an F1 score in a range between 0.86 and 0.97 across all languages for `butnot`, `comparatives`, `drather`, `prefer`, and `andtoo` constructions which made up approximately 42.3% of the test dataset. These scores are about what we expected from performance on the validation set.

`except`, `particular`, and `unlike` constructions show stark differences in performance between languages. In all three cases, the French model achieves much better results than the English and Italian ones, which will be discussed in more detail later. Furthermore, the models performed poorly on `type` constructions across all languages, only overshadowed by scores close to 0 for generalizations.

4.1.1 `except`

Sentences containing `except` constructions made up approximately 12.9% of the test data. Especially the Italian model seemed to have trouble classifying exceptions correctly. For example, it gives the sentence *Adoro le verdure, eccetto le carote.*[2] the label "0" even though it is a semantically flawless Italian sentence. The problem seems to be the recall rather than precision. While a score of 0.82 is not much lower than the precision of the previously mentioned constructions, the recall score of 0.31 shows that the model could not accurately predict the taxonomic relations in a sentence containing an exception. This observation extends, if less prominently, to English and French.

4.1.2 `particular`

`particular` sentences, which make up 13% of the test data, show a striking difference in performance between languages. The precision ranges from 0.39 for the English model to 0.92 for the French model.

4.1.3 `unlike`

Sentences containing `unlike` constructions (9% of the test data) were still classified correctly relatively often by the French model, despite a recall score of 0.63. The Italian and English models performed much worse due to low recall. Interestingly, the English model has an almost-perfect precision of 0.99, while the Italian model only reached 0.49, which shows the difference between the models again.

[2] *I love vegetables, except for carrots.*

	total	It	Fr	En
MSE	3.72	2.88	5.29	2.97
RMSE	1.93	1.70	2.30	1.72
RHO	0.19	0.19	-0.01	0.06

Table 3: Results for subtask 2. All scores are micro-averaged over different constructions. The total scores are also micro-averaged over the languages.

4.1.4 `type`

About 13% of the test sentences contained `type` constructions like *I like fruits, an interesting type of lemon.* There is not much to say about them other than that the models' performance on them was terrible. F1 measures range from 0.06 to 0.11 with very low recall (0.25 to 0.47) and even worse precision (close to 0).

4.1.5 `generally`

Possibly the most surprising result is the utter confidence with which the models misclassified generalizations. Across languages, all evaluation metrics are below 0.05 for sentences labeled `generally`.

4.2 Subtask 2

The top-performing models reached a Spearman correlation of 0.81, 0.84, and 0.76 in Italian, French, and English, respectively, yielding an overall score of 0.80. The KaMiKla models performed much worse and ranked in 11th place with a global rank of 0.08. Table 3 shows an overview of the evaluation of the second subtask. Surprisingly, the n-gram-based regression model serving as a baseline ranked in 4th place with correlations of 0.34, 0.32, 0.27 in Italian, French, and English, respectively, outperforming many submissions, including the one present in this paper.

Similar to the first subtask, we analyzed these results grouped by the used construction. Appendix D contains the detailed analysis. The metrics considered are the mean squared error (MSE), the root mean squared error (RMSE), and Spearman correlation (RHO). The constructions of the second subtask are a subset of the ones used for the first one, with the `generally` label changed to `ingeneral`. Figure 3 visualizes the root mean squared errors of the second subtask grouped by construction.

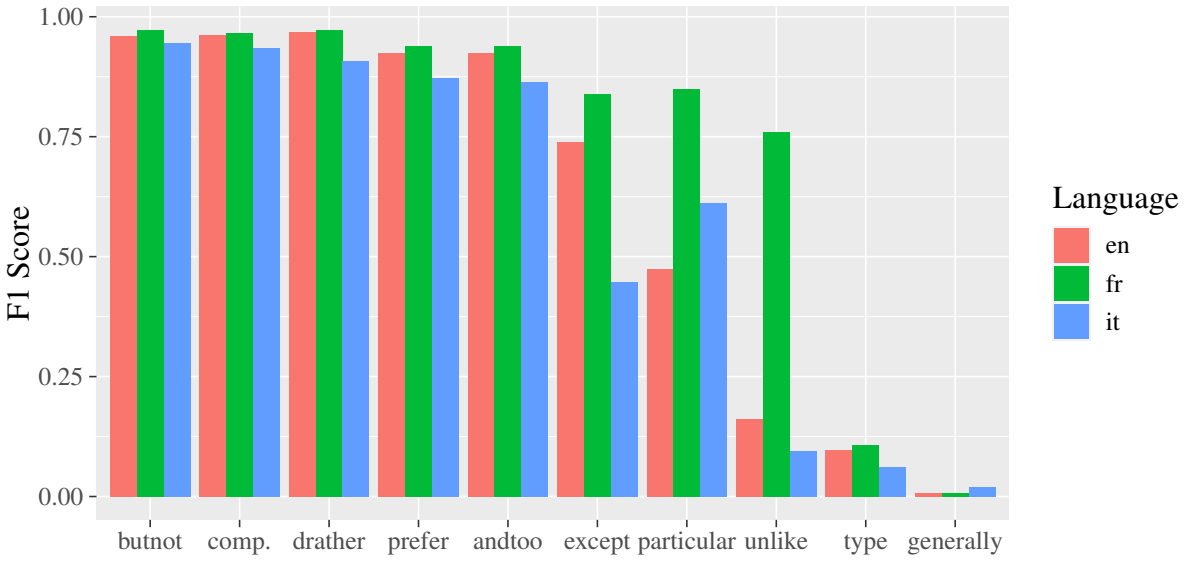

Figure 2: An overview over the F1 scores the KaMiKla models reached in subtask 1, ordered by score.

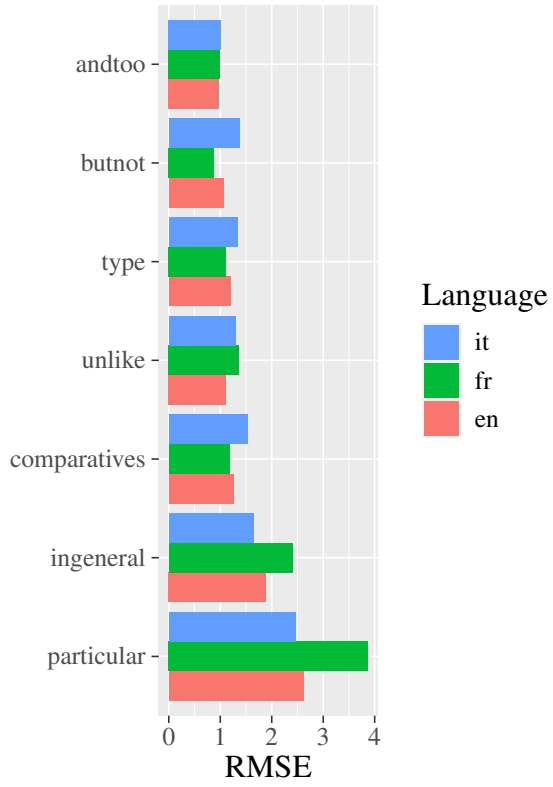

Figure 3: An overview over the root mean squared error scores the KaMiKla models reached in subtask 2, ordered by score.

The difference in performance between constructions is evident here as well. Interestingly, the French regression model seemed to have more trouble than its Italian and English counterparts, while the French classification model outperformed the other languages.

Another interesting comparison is that of the constructions. While the classification models failed on `type` sentences, those are some of the most successful sentences in subtask 2; this indicates that the bad scores on some constructions are due to fine-tuning and not some inherent difficulty the BERT models have with understanding them. On the other hand, this might be a side effect of the models not performing very well in general.

5 Discussion

Our models performed considerably worse on the test dataset than on the validation set, and these differences vary across languages and constructions. While we can't determine for sure where this significant drop in performance comes from, there are some theories worth investigating.

Of course, one challenge (especially of the second part) of the task is the scarcity of data. Subtask 1 contains 5,837 sentences in the training data per language; and 14,560 sentences in the test set. For subtask 2, there are 524 training sentences and 1,009 test instances in each language. Possibly, the regression models performed unsatisfyingly simply because there wasn't enough training data avail-

able.

We also wanted to investigate the inherent differences between train and test data. Figure 4 compares the sentence lengths of the train and test set, grouped by language. There are huge differences between the languages and, perhaps more importantly, between the train and test sets of the three languages. The French training data contained sentences of similar length to those in the test data (mean of 8.94 for the test and 9.28 for the training set), and those sentences were generally longer than those of the other two languages. The data from the Italian test set seems to contain sentences of more varied lengths than the training set.

The data of the second subtask show a similar discrepancy between training and test sentences. The sentences from the test data are longer on average than those from the training data in all languages. The lengths also seem to be more variable in the test set, possibly due to the higher number of sentences. The gap between the input sentences could have led to finetuning not remarkably improving the performance of the models.

We furthermore looked at the distribution of different types of constructions in more detail. There were no construction labels for the training data, so we trained a simple Naive Bayes classifier on the tf-idf-transformed test sentences in Python using pandas (pandas development team, 2022) and scikit-learn (Pedregosa et al., 2011). Because of the simple structure of the input sentences and a cross-validation score of 100%, we will assume the construction labels to be accurate in further discussion.

Figures 5 and 6 compare the distributions of the construction labels in train and test sets of subtask 1 and 2, respectively. There are once again huge differences between training and test set. Notably, the training data does not contain a single `unlike` sentence. Despite that, the models did not necessarily perform worse on this type of construction. In total, it does not seem like the distribution of construction types in the training data influenced model performance much at all. There are instances of models performing inadequately on frequent constructions in the training data, like `type` constructions in the first subtask. However, `drather` constructions were often classified correctly despite the scarcity of training sentences.

6 Conclusion

In this work, we discussed an approach to modeling taxonomic relationships using pre-trained language models, namely AlBERTo (Polignano et al., 2019), BERT (Devlin et al., 2019), and CamemBERT (Martin et al., 2020) in the context of the SemEval Task 3 of the year 2022. The KaMiKla group participated in both the classification and the regression subtask. While the performance of the models was overall unsatisfying, further analysis revealed that the type of taxonomic relation that the words in a given sentence severely affected how well the models did. While the reason for this remains unclear, it might be interesting to tailor the finetuning of the BERT-based model to specific constructions or combine it with a classifier that classified the input sentences according to the type of taxonomic relation.

Acknowledgements

We want to thank Çağri Çöltekin for his technical and topical assistance throughout this project.

References

Michael Clarke. 2012. 4 - the digital revolution. In Robert Campbell, Ed Pentz, and Ian Borthwick, editors, *Academic and Professional Publishing*, pages 79–98. Chandos Publishing.

David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. 2019. *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.

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

Ingo Feinerer, Kurt Hornik, and David Meyer. 2008. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54.

Paul Kay. 1971. Taxonomy and semantic contrast. *Language*, 47:866–887.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Max Kuhn. 2021. *caret: Classification and Regression Training*. R package version 6.0-90.

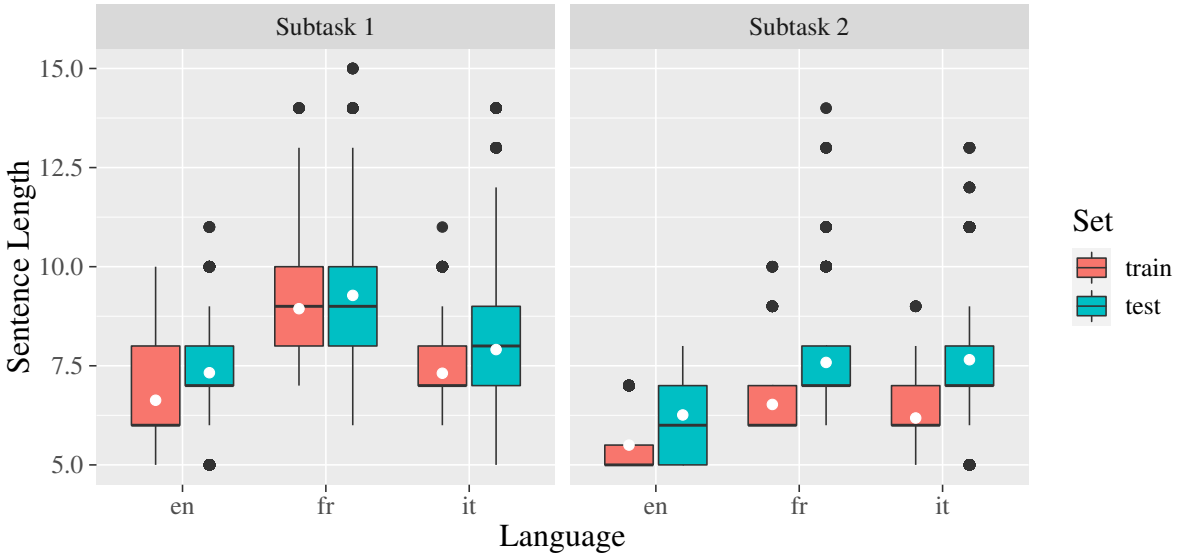

Figure 4: A comparison of the sentence lengths of train and tests sets in the different languages for both subtasks. The mean is visible in white. The difference seems to be less pronounced in the French data of subtask 1, which might have contributed to the better scores. In subtask 2, the sentences seem to be generally longer in the test data than the training data in all languages.

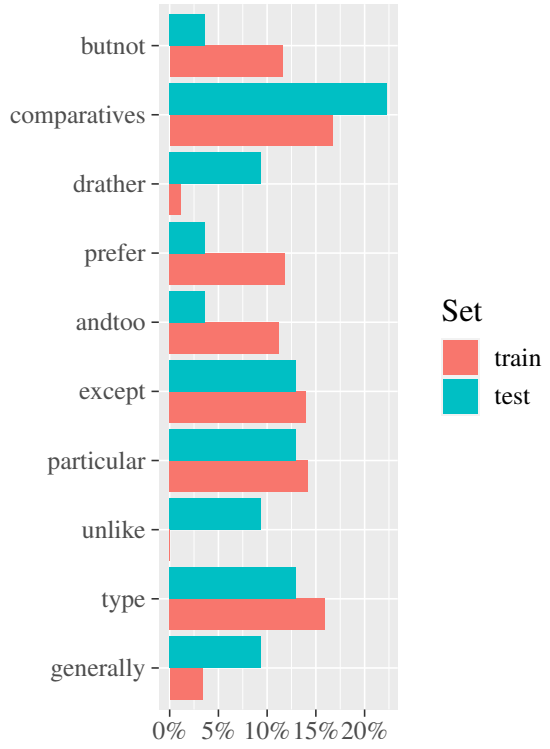

Figure 5: A comparison of the distribution of different constructions in train and test set of subtask 1, ordered by their score in the classification task. The x-axis shows the percentage of this label in one of the datasets.

Figure 6: A comparison of the distribution of different constructions in train and test set of subtask 2, ordered by their score in the regression task. The x-axis shows the percentage of this label in one of the datasets.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. Published as a conference paper at ICLR 2019.

Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413, Austin, Texas. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen, Denmark. Association for Computational Linguistics.

The pandas development team. 2022. pandas-dev/pandas: Pandas 1.4.1.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. ALBERTO: Italian BERT language understanding model for NLP challenging tasks based on tweets. pages 1–6.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Charlie Sharpsteen and Cameron Bracken. 2020. *tikzDevice: R Graphics Output in LaTeX Format*. R package version 0.12.3.1.

Hadley Wickham. 2011. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Hadley Wickham. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2021. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.6.

Hadley Wickham and Jim Hester. 2020. *readr: Read Rectangular Text Data*. R package version 1.4.0.

Roberto Zamparelli, Shammur A. Chowdhury, Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Arid Hasan, and Giulia Venturi. 2022. SemEval-2022 Task3 (PreTENS): Evaluating neural networks on presuppositional semantic knowledge. In *Proceeding of SEMEVAL 2022*.

A Hyperparameters For Subtask 1

Table 4 shows the hyperparameters used in subtask 1 for the different languages.

B Results For Subtask 1

Tables 5, 6, and 7 show the performance metrics of the BERT-based models in the classification task. There are considerable differences between constructions as well as between languages.

C Hyperparameters For Subtask 2

Table 8 shows the hyperparameters used in subtask 2 for the different languages.

D Results For Subtask 2

Tables 9, 10, and 11 show the metrics for the regression task.

	Batch Size	Learning Rate	Max. Len. Sent.	Patience	Hidden Layers	
					Number	Size
English	32	1×10^{-5}	15	5	2	512
French	32	1×10^{-5}	20	5	2	512
Italian	32	1×10^{-5}	15	5	2	512

Table 4: Hyperparameters used in subtask 1.

	total	***andtoo***	***butnot***	***comp.***	***drather***	***except***	***generally***	***particular***	***prefer***	***type***	***unlike***
Precision	0.74	0.86	0.93	0.93	0.94	0.85	0.01	0.39	0.88	0.06	0.99
Recall	0.80	1.00	0.99	0.99	1.00	0.65	0.01	0.62	0.98	0.40	0.09
F1	0.77	0.93	0.96	0.96	0.97	0.74	0.01	0.47	0.93	0.10	0.16

Table 5: Metrics for the English sentences in subtask 1, grouped by construction.

	total	***andtoo***	***butnot***	***comp.***	***drather***	***except***	***generally***	***particular***	***prefer***	***type***	***unlike***
Precision	0.77	0.90	0.95	0.94	0.95	0.94	0.01	0.92	0.90	0.06	0.96
Recall	0.89	0.98	1.00	0.99	1.00	0.76	0.01	0.79	0.98	0.47	0.63
F1	0.83	0.94	0.97	0.97	0.97	0.84	0.01	0.85	0.94	0.11	0.76

Table 6: Metrics for the French sentences in subtask 1, grouped by construction.

	total	***andtoo***	***butnot***	***comp.***	***drather***	***except***	***generally***	***particular***	***prefer***	***type***	***unlike***
Precision	0.73	0.83	0.92	0.92	0.94	0.82	0.03	0.74	0.85	0.03	0.49
Recall	0.73	0.90	0.97	0.95	0.87	0.31	0.02	0.52	0.89	0.25	0.05
F1	0.73	0.86	0.95	0.93	0.91	0.45	0.02	0.61	0.87	0.06	0.09

Table 7: Metrics for the Italian sentences in subtask 1, grouped by construction.

	Batch Size	Learning Rate	Max. Len. Sent.	Patience	Hidden Layers	
					Number	Size
English	32	1×10^{-5}	15	5	1	16
French	32	1×10^{-5}	20	5	1	16
Italian	32	1×10^{-5}	15	5	1	16

Table 8: Hyperparameters used in subtask 2.

	total	***andtoo***	***butnot***	***comparatives***	***ingeneral***	***particular***	***type***	***unlike***
MSE	2.97	0.96	1.13	1.59	3.57	6.83	1.45	1.24
RMSE	1.72	0.98	1.06	1.26	1.89	2.61	1.20	1.11
RHO	0.06	-0.32	0.04	0.07	-0.28	-0.25	0.26	0.18

Table 9: Metrics for the English sentences in subtask 2, grouped by construction.

	total	***andtoo***	***butnot***	***comparatives***	***ingeneral***	***particular***	***type***	***unlike***
MSE	5.29	1.00	0.77	1.39	5.77	14.99	1.24	1.87
RMSE	2.30	1.00	0.88	1.18	2.40	3.87	1.11	1.37
RHO	-0.01	0.43	0.19	0.35	-0.03	-0.29	0.25	0.38

Table 10: Metrics for the French sentences in subtask 2, grouped by construction.

	total	***andtoo***	***butnot***	***comparatives***	***ingeneral***	***particular***	***type***	***unlike***
MSE	2.88	1.03	1.90	2.32	2.71	6.05	1.79	1.69
RMSE	1.70	1.02	1.38	1.52	1.65	2.46	1.34	1.30
RHO	0.19	-0.21	-0.05	-0.28	-0.07	-0.00	-0.18	0.34

Table 11: Metrics for the French sentences in subtask 2, grouped by construction.