SemEval-2022 Task 3: PreTENS – Evaluating Neural Networks on Presuppositional Semantic Knowledge

Roberto Zamparelli[1], Shammur A Chowdhury[2], Dominique Brunato[3], Cristiano Chesi[4], Felice Dell'Orletta[3], Arid Hasan[5], Giulia Venturi[3]
[1]CIMeC, University of Trento, Rovereto Italy;
[2]Qatar Computing Research Institute, HBKU, Qatar; [3]ILC-CNR, Pisa, Italy;
[4]NETS-IUSS, Pavia, Italy; [5]Daffodil International University, Dhaka, Bangladesh
roberto.zamparelli@unitn.it, shchowdhury@hbku.edu.qa,
cristiano.chesi@iusspavia.it, arid.cse0325.c@diu.edu.bd, [name.surname]@ilc.cnr.it

Abstract

We report the results of the SemEval 2022 Task 3, PreTENS, on evaluation the acceptability of simple sentences containing constructions whose two arguments are presupposed to be or not to be in an ordered taxonomic relation. The task featured two sub-tasks articulated as: *(i)* binary prediction task and *(ii)* regression task, predicting the acceptability in a continuous scale. The sentences were artificially generated in three languages (English, Italian and French). 21 systems, with 8 system papers were submitted for the task, all based on various types of fine-tuned transformer systems, often with ensemble methods and various data augmentation techniques. The best systems reached an F1-macro score of 94.49 (sub-task1) and a Spearman correlation coefficient of 0.80 (sub-task2), with interesting variations in specific constructions and/or languages.

1 Introduction

A growing body of literature on the cognitive side of computational linguistics has tried to probe the ability of language models to recognize deviant linguistic structures. Recognizing linguistic ill-formedness requires some degree of metalinguistic awareness in adult humans (i.e. the ability to say not just that there is 'something wrong' in a sentence, but also where the problem lies or how the sentence could be improved), and it is not clear whether and to what extent artificial systems can induce this type of knowledge purely from exposure to raw linguistic data (Linzen and Baroni, 2020). Most of the previous investigations on this topic have focused on phenomena that are purely syntactic (agreement, Gulordava et al. 2018; dislocated arguments with island effects, Wilcox et al. (2018); Warstadt et al. (2019); Chowdhury and Zamparelli (2018), clause embedding, Futrell et al. 2019, etc.) or at the syntax/semantics interface (negative-polarity items, Jumelet and Hupkes 2018; argument structure, quantifier restrictions, Warstadt et al. 2019), mostly using LSTM architectures (but see Tran et al. 2018; Ettinger 2020). The fact that many purely semantic effects result in the (non) availability of certain readings (e.g. the scope of a quantifier over a higher negation, in "I didn't see some people") makes it of course harder to translated them into computationally testable paradigms.

The task we describe in this paper focuses on an area of purely semantic competence that, to the best of our knowledge, is unexplored in the computational literature, and one which gives rise to a robust intuition of deviance, triggered by the failure of a type of **presupposition**: the requirement for two nominal arguments to be (or not be) in an (ordered) **taxonomic relation**. These presuppositions are introduced by a wide variety of constructions, such as comparatives (1a), coordinations, verbs like *prefer*, modifiers headed by *type* or *except* (1b) etc. (see Table 1 for the full list).

(1) a. I hate guns more than {*weapons / social media}.
b. I like dogs, except {*birds / greyhounds}

Distinguishing the deviant from the acceptable cases requires the ability to (i) detect taxonomic relations and (ii) recognize those constructions which place restrictions on them. The first point is of course crucial for most tasks in Natural Language Inference (NLI) — an active and fast-growing field in the NLP community, with various datasets and benchmarks (e.g. GLUE Wang et al. 2018, SuperGLUE Wang et al. 2019). NLI datasets, however, normally assume that felicity conditions are satisfied. The present dataset, which we call PreTENS, takes a step back and aims to verify if a computational model can detect when this presupposition fails.

The task requires world knowledge, commonsense knowledge (in the sense discussed in Storks et al. 2019), but also linguistic knowledge, to catch

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 228 - 238

Construction (Variants)	Example	Presup.
EXCEPT (2)	I like [$_{A1}$ dogs] except [$_{A2}$ {*cats / pugs / *animals}]	**A1>A2**
PARTICULAR (2)	I like [$_{A1}$ dogs], and in particular [$_{A2}$ {*animals / *cats / pugs}].	**A1>A2**
IN GENERAL	I like [$_{A1}$ dogs], and [$_{A2}$ {animals / *cats / *pugs}] in general.	**A1<A2**
GENERALLY	I like [$_{A1}$ dogs], and more generally [$_{A2}$ {animals / *cats / *pugs}].	**A1<A2**
TYPE (2)	I like [$_{A1}$ dogs], an interesting type of [$_{A2}$ {animal / *cat / *pug}].	**A1<A2**
AND-TOO	I like [$_{A1}$ dogs], and [$_{A2}$ {cats / *pugs / *animals}] too.	**A1⪅A2**
COMPAR. (3)	I like [$_{A1}$ dogs] more than [$_{A2}$ {cats / *pugs / *animals}]	**A1⪅A2**
DRATHER	I would rather have [$_{A1}$ dogs] than [$_{A2}$ {cats / *pugs / *animals}]	**A1⪅A2**
PREFER	I don't like [$_{A1}$ dogs], I prefer [$_{A2}$ {cats / *pugs / *animals}]	**A1⪅A2**
UNLIKE	Unlike [$_{A1}$ dogs], [$_{A2}$ {*animals / cats / *pugs}] are often mentioned in this text .	**A1⪅A2**
BUT-NOT	I like [$_{A1}$ dogs], but not [$_{A2}$ {*animals / cats / pugs}]	**A1⪅/>A2**

Table 1: Distribution of taxonomic constructions and their presuppositions. ⪅ indicates no overlap; (*n*) indicates n variants on the construction (e.g. COMPAR. contains samples of majority, minority and equality comparatives). The BUT-NOT case is probably ambiguous, with one meaning close to EXCEPT; the same applies to GENERALLY, which draw uncertain results in the human evaluation and was excluded from sub-task 2 in favour of IN GENERAL.

the requirement of the specific presupposition-inducing constructions. In this respect, the present task is closer in spirit to SemEval-2020 task 4, subtask A, on the validation of sentences for commonsense (Wang et al., 2020), than to SemEval 2016 task 3, where participants had to extract and identify the taxonomic relationships between two terms (Bordea et al., 2016).

Besides NLI practitioners, the task could be relevant for researchers interested in the potential of NNs as cognitive/linguistic models (see e.g. Warstadt et al. 2019). We believe that it is also a potentially useful addition to the toolbox of probes used to understand the inner working of current language models.

2 Dataset and Task description

2.1 Composition

The PreTENS contains 21,765 artificial sentences with 2-argument relations filled by nominals (the **argument nouns**). The sentences were designed to follow or flout the presuppositions that (i) the argument nouns should or should not be in a taxonomic relation (i.e. one a subset of the other: *dogs* < *animals*) and (ii) when a taxonomic relation was required, the order should be a specific one. (ii) differentiates *I like dogs, and in particular pitbulls* from **I like pitbulls, and in particular dogs*). The list of constructions used is in Table 1.

The data for this task was programmatically generated from a human-verified template, yielding sets of sentences that are extremely similar across constructions. The argument nouns (A1 and A2 in Table 1) are taken from the following semantic categories: *dogs, birds, animals, cars, motorcycles, cutlery, clothes, trees, plastics, furniture, wine, animals, sports, music, vegetables, fruits, pork-based food, desserts, seafood, apartments, movies, jewelry, pets, rain, nature, senses, emotions, books, workers* and *scientists*, and repeat across constructions. The elements not in taxonomic relations were chosen to maximize the plausibility of comparison (e.g. *dogs* if the semantic category was *birds*) and the verbs were chosen to be as semantically neutral as possible (often *like* or *have*, but e.g. *trust* in the semantic category of *senses*). The English template file was created and revised (using dictionaries and Wordnet) by the task proponents, all expert linguists, and double-checked by a native speaker.

PreTENS is a simplified, no-repetiton subset of a larger dataset, DuckRabbit, which also contains 5 semantic categories (*countries, cities, painters, politicians, actors*) examplified by well-known proper names (e.g. *Paris, Picasso, Obama*), which we decided not to use for the PreTENS task. The full DuckRabbit dataset (55,296 items) is arranged in a way that systematically tests all the possible orders of pairs of argument nouns taken from a supercategory, a subcategory in the same taxonomic domain and a distractor (non taxonomically ordered with either, e.g. <*birds, parrots, dogs*>). This arrangement, however, creates a large number of repeated entries.

The fixed nature of the patterns used allowed us to propose the dataset in three languages (English, Italian and French), where the French and Italian versions are slightly adapted translations of the English dataset.[1] Adding more languages would

[1] A key difference was that the English bare plurals used in generic sentences were replaced by NPs introduced by definite

be relatively straightforward. The template and the scripts used to generate the data are publicly available under a CC BY 3.0 "Attribution" license.[2] As far as we can tell, the contents do not raise any issue w.r.t. ethics or privacy.

2.2 Definition of the Task

The task was articulated into the two sub-tasks:

- a **binary classification task** (hereafter referred to as *sub-task1*), which consisted in predicting the acceptability label assigned to each sentence of the test set on the basis of a theoretical linguistic model;
- a **regression task** (hereafter *sub-task2*), which consisted in predicting the average score on a 7 point Likert-scale assigned by human annotators to a subset of data evaluated via crowdsourcing (see Section 2.3).

For each task and each language, the dataset was split into training and test sets. The classification task was composed of 5,838 training samples and 14,560 testing samples; the regression task, of 524 sentences in training and 1,009 in test. Table 2 reports the internal composition of the training and test dataset of each sub-task. As it can be seen, not all the constructions contained in the test were provided in the training set. This choice was deliberate, to test the generalization abilities of the systems across unseen constructions. The sentences in training data were independently randomly ordered in the three languages, to discourage mapping the results obtained in one language to sentences with the same ID in the other languages.

2.3 Annotation with human judgments

The dataset released for *Sub-task2* is composed by a subset of 1,533 sentences taken from the whole dataset, corresponding to about 5% of the total and representative of the patterns considered, which were judged by human annotators via a crowdsourcing campaign.

The purpose of this evaluation was two-fold: (i) to provide a bottom-up assessment of the quality of the linguistic categories that informed the creation of the dataset templates; (ii) to obtain more fine-grained judgments of semantic acceptability in the form of gradual, rather than categorical, scores.

Constructions	sub-task1		sub-task2	
	Training	**Test**	**Training**	**Test**
and-too	835	525	131	88
but-not	831	526	131	88
comparatives	835	3,245	131	88
drather	–	1,360	–	–
except	831	1,887	–	–
in general	–	–	–	219
generally	–	1,360	–	–
particular	835	1,885	–	219
prefer	835	525	–	–
type	835	1,887	131	88
unlike	–	1,360	–	219
TOTAL	**5,838**	**14,560**	**524**	**1,009**

Table 2: Distribution of taxonomic constructions in terms of number of sentences in the dataset.

The annotation was performed through the Prolific[3] platform. Specifically, for each language the annotation process was split into different tasks, each one consisting in the annotation of about 150 randomly mixed sentences for the typologies reported in Table 1. For all tasks, we recruited 12 native speakers, who were asked to read each sentence and answer the following question:

> *How acceptable is this sentence from 1 (completely unacceptable) to 7 (completely acceptable)?'*

As an example, we report below two sentences (with corresponding average score) from the English portion of the annotated dataset, which were rated as very poorly and very highly acceptable:

> *I like politicians, an interesting type of farmer* (1.42)
>
> *I like governors, an interesting type of politician* (6.16)

Table 3 provides the average value (μ) and standard deviation (σ) of acceptability labels for the whole dataset (first row) and for sentences classified according to the various constructions. As it can be noted, French sentences were evaluated on average as more acceptable than Italian and English ones but with a slightly higher standard deviation. While for all languages the maximum average score on the Likert scale was obtained by very few sentences (i.e. only one sentence for English and French and four for Italian), the number

determiners in Italian and French. This makes the latter sub-datasets systematically longer. In addition, certain English nouns required compounds or N+PPs to be rendered in the other languages.

[2]Github Repository: https://github.com/shammur/SemEval2022Task3
Task Website: https://sites.google.com/view/semeval2022-pretens/

[3]www.prolific.co

	ENG		ITA		FRE	
	μ	σ	μ	σ	μ	σ
All_sents	3.89	1.61	3.75	1.73	4.05	1.85
and-too	4.83	0.92	4.98	0.97	5.29	0.95
but-not	4.59	1.08	3.94	1.06	5.07	0.94
comparatives	4.91	1.93	5.05	1.42	5.18	1.41
in general	3.64	1.17	3.28	1.29	3.78	1.06
particular	2.54	1.48	2.36	1.46	2.28	1.50
type	2.13	1.45	1.89	1.44	1.80	1.40
unlike	4.56	0.84	4.74	1.15	4.93	1.93

Table 3: Statistics about the distribution of human judgments in the dataset collected for sub-task2. μ = average judgment; σ = standard deviation.

of sentences rated with the lowest score is higher for Italian and French (i.e. 42 and 45 respectively) than for English (i.e. 7). If we focus on the distribution of judgments across the distinct constructions, we observe that examples containing the TYPE construction were perceived on average as the less acceptable ones for all languages. Conversely, sentences belonging to the AND TOO and COMPARATIVES categories obtained the highest acceptability scores.

In order to see how consistent was the human perception of semantic acceptability across languages, we computed the Pearson's r between the average scores assigned to the whole set of sentences for each pair of languages. The correlation scores were very high, with the highest scores obtained between sentences in French and Italian (i.e. 0.86), followed by English and French (i.e. 0.80), and, lastly, by English and Italian (i.e. 0.77)[4].

Finally, an additional outcome that we want to highlight here is the strong connection between the theoretically-driven and the human-based semantic acceptability label, which was assessed by calculating the Spearman's rank correlation coefficient between the average human scores and the binary acceptability labels attributed to the same set of sentences, for all languages. In this case, too, we found a very high correlation, although weaker in English (ρ=.73) than in Italian and, especially, French (ρ =.78 and .83, respectively).

3 Shared Task Organisation

Shared Task Phases We ran the shared Task 3 in two phases. In the first phase, we released the baseline pipeline, along with the cross-validation results on the official training set and introduced the participants to the aforementioned task evaluation measure.

[4]All correlations are significant with p value$<$ 0.01.

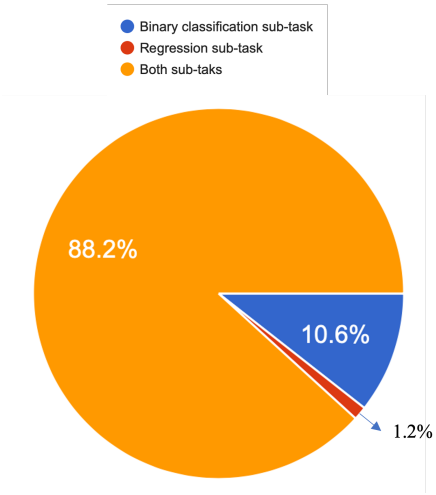

Figure 1: Statistics of participants' interest on Tasks based on initial registration.

Task	sub-task1	sub-task2
Team Participated	21	17
Total System Submissions	134	110
Total Accepted Submissions	108	84

Table 4: Statistics on participation

The second phase – the main Evaluation Phase – was conducted using codalab platforms for both sub-task1[5] and sub-task2.[6] During this phase, the participants were provided with the test sets and were allowed to submit their predictions to the system. The number of submissions of each participant was limited to three, but the participant could choose among them which runs/submission they want to display in the leader-board. During the evaluation phase, the leader-board was visible to all the participants.

Baselines For each sub-task a separate baseline were defined: *i)* for Sub-task1 – the binary classification sub-task, a Linear Support Vector classifier using n-grams (up to three) as input features was used, and for the *ii)* Sub-task2 – the regression sub-task, a baseline using a Linear Support Vector regressor with the same n-grams features was provided.

We provided the starter code to all the participants, along with different cross-validation configurations that we encouraged participants to use to validate their methodology. Moreover, we provided

[5]https://codalab.lisn.upsaclay.fr/competitions/1292

[6]https://codalab.lisn.upsaclay.fr/competitions/1290

information on how the performance in the validation task translated in the official test split (applying the baseline methods to the official test-set yielded results 10-20% lower than with the training set). We highlighted the importance of achieving maximal syntactic generality on this task and test different cross-validation configurations on the training set.

Official Evaluation Metrics Given the differences in the nature of the sub-tasks output, we defined two different sets of evaluation metrics. For sub-task1, systems are evaluated with respect to binary and macro F-measure. These measures were evaluated per languages, and the final ranking was based on the global ranking of each participant, calculated by averaging the macro F-measure score from all the three languages (this provided an incentive to give results for all languages). In addition to the official measure, we also gave the participants their precision and recall scores, per language.

As for sub-task2, a Spearman's rank correlation coefficient (ρ) between the task participants' scores and the test set scores was computed. To be consistent with sub-task1, the global ranking of this task was calculated by averaging the position of the participant's ρ per language.

At the end of the competition, we provided the participants with packages containing the results for each of their submissions, and publicly updated the leader-board with ranks listing all teams who competed in each sub-task.

Participation The task attracted nearly 83 teams. Among them, 43 teams actually registered for the evaluation phase; 21 teams (sub-task1) and 17 teams (sub-task2) submitted their system's predictions. The detailed statistics are shown in Table 4.

4 Participating Systems

We received six system description papers for both sub-tasks, plus two papers by teams that participated only in sub-task1, for a total of eight papers. As it can be observed in the following summaries of the approaches proposed, there were several points of methodological similarity, but also interesting differences. Many teams experimented data augmentation techniques devised to overcome the limited amount of training data, in particular for the solution of sub-task2. These techniques ranged from the use of external resources to the generation of new sentences (Zhou et al. (2022), Sarhan et al. (2022) and van den Berg et al. (2022)), to the automatic translations across the three languages considered (Sarhan et al. (2022) and Zhou et al. (2022)), to mapping the Likert scale results to the binary values. This was the strategy used by the first two teams classified (Xia et al. (2022) and Li et al. (2022)) according to the global scores.

As we can see by the description of the participating systems (see Section 4.1), the majority of teams chose monolingual instead of multilingual models, especially in the resolution of sub-task1. The exceptions are represented by Li et al. (2022), who obtained the second position in the global ranking of both sub-tasks, by Aziz et al. (2022), and by Sarhan et al. (2022) who (in the resolution of subtask 2 only), used the multilingual version of the Universal Sentence Encoder, since it yielded better performance than the monolingual one. Interestingly enough, the top-ranked team in both sub-tasks (Xia et al. (2022)) found that, for all languages, the monolingual DeBERTa-v3 models always outperformed the multilingual version.

A further approach shared by the participating teams is represented by the adoption of ensemble methods. Two main ensemble strategies were suggested. In the first one, the training data used to fine-tune the adopted model was split, obtaining different models, each with its training and validation sets. This was the case with the second-ranked system (Li et al. (2022) and with Vetter et al. (2022), but only in sub-task2. A second main approach used the fusion of the acceptability scores predicted by two different models. As described in the following subsection, Aziz et al. (2022) combined the scores predicted by XLM-RoBERTa (Conneau et al., 2019) and mBERT (Devlin et al., 2019), while Zhou et al. (2022) merged the predictions made by ERNIE-M and DeBERTa-v3 (only in subtask1).

4.1 Individual System Descriptions

Xia et al. (2022) (model LingJing), tackling subtask1, experimented with different strategies to fine-tune DeBERTa-v3 (He et al., 2021), which ended up outperforming both the PreTENS baseline and three new baselines introduced by the authors, including a multilingual version of DeBERTa, i.e. mDeBERTa model. These strategies included the augmentation of the original training set with trans-

Team/user name	Global score	Rank	ENG	Rank	FRE	Rank	ITA	Rank
LingJing⋆	94.49	1	97.17	1	93.24	1	93.05	1
HW-TSC⋆	92.80	2	93.04	5	93.01	2	92.34	2
UU-TAX⋆	91.57	3	93.08	4	89.53	4	92.12	3
CSECU-DSG⋆	91.12	4	91.51	7	90.73	3	91.11	4
piano	90.74	5	97.12	2	86.14	11	88.95	6
SPDB Innovation Lab⋆	89.58	6	94.55	3	87.28	7	86.90	8
bpc	89.09	7	91.36	8	88.32	6	87.59	7
weijiyao	88.78	8	92.15	6	87.28	8	86.90	9
ddd7788	86.68	9	80.44	13	88.86	5	90.75	5
cnxupupup	86.68	10	86.88	9	86.39	10	86.76	10
MaChAmp	86.42	11	86.58	10	86.52	9	86.17	12
aidenqiu	86.30	12	86.29	11	86.09	12	86.51	11
UoR-NCL⋆	80.32	13	77.23	16	80.08	14	83.65	13
RUG-1-pegasussers⋆	79.56	14	80.31	14	79.71	15	78.64	14
KaMiKla⋆	77.99	15	77.21	17	82.34	13	74.40	15
Huawei-zhangmin	71.80	16	78.54	15	65.77	18	71.08	16
RCLN	70.54	17	73.02	18	75.73	16	62.86	17
BASELINE	67.39	18	70.47	19	72.13	17	59.59	18
Jan/Jasper/Boris	27.26	19	81.76	12	–	–	–	–
folkertleistra	22.64	20	67.92	20	–	–	–	–
RUG-3	19.95	21	59.85	21	–	–	–	–

Table 5: Sub-task 1 results for each team/user ordered by overall F1-Macro along with micro-averages for each language. Team/user names marked with ⋆ have submitted their system description.

lations from the three languages, adversarial training and Child-Tuning (Xu et al., 2021). In addition, the authors performed experiments with different compositions of the original training, i.e. mixing the data for the three languages, and fine-tuning in one language, then expanding to the others. Each strategy achieved different results for each language. Due to the small size of the training set of sub-task2, the team transferred the knowledge of the classification model to the regression task, in terms of the model's parameters and in the idea of mapping the Likert scale results to the binary values.

Li et al. (2022) (HW-TSC), addressing both subtasks1 and 2, developed an ensemble classification and regression model by fine-tuning the multilingual XLM-RoBERTa model (Conneau et al., 2019) on different splits of the training data. To this end, they added a language tag to each training sentence with the same id across the three languages and divided the data in different folds to prevent the model from learning the translation information. To address the small size of the sub-task2 training data, they devised a data augmentation strategy to transform the binary values into the scalar human judgments.

Sarhan et al. (2022) (UU-TAX), experimented with different Neural Language Models and diverse training data compositions to test their generalization abilities against the PreTENS tasks. For subtask1, their best performing model is represented by ELECTRA (Clark et al., 2020), which was fine-tuned using a two-stage strategy to augment the original training data. Firstly, the authors generated new sentences by making modifications to the original sentences using BERT-base (Devlin et al., 2019) to obtain the embeddings of the modified words. Secondly, for each language l, the original sentences of the other two languages were translated into l using the Google Translate API. For sub-task2, the best model uses the multilingual version of the Universal Sentence Encoder (Yang et al., 2020) followed by a different type of classifier for each language.

Aziz et al. (2022) (CSECU-DSG), for both subtasks, exploited an ensemble method of two multilingual Transformers, i.e. XLM-RoBERTa (Conneau et al., 2019) and mBERT (Devlin et al., 2019), which were fine-tuned with the PreTENS datasets. To enhance the performance of each individual model, the authors fused the predicted probability scores of the two models by computing their weighted arithmetic mean.

Vetter et al. (2022) (KaMiKla), for both subtasks, used monolingual versions of the BERT model (Devlin et al., 2019), i.e. BERT base for English, AlBERTo (Polignano et al., 2019) for Italian and CamemBERT (Martin et al., 2020) for French. For sub-task1, the authors fine-tuned the models on the distributed training data, while for sub-task2,

Team/user name	Global score	Rank	ENG	Rank	FRE	Rank	ITA	Rank
LingJing⋆	0.802	1	0.758	1	0.841	1	0.807	1
HW-TSC⋆	0.757	2	0.706	2	0.805	2	0.759	2
Huawei-zhangmin	0.669	3	0.636	3	0.74	3	0.631	3
BASELINE	0.309	4	0.265	6	0.317	4	0.344	4
UU-TAX⋆	0.221	5	0.478	4	-0.062	15	0.246	5
daydayemo	0.206	6	0.212	8	0.284	5	0.121	9
aidenqiu	0.205	7	0.211	9	0.284	6	0.121	10
CSECU-DSG⋆	0.16	8	0.191	10	0.081	8	0.207	6
RCLN	0.139	9	0.418	5	-0.005	9	0.006	14
folkertleistra	0.123	10	0.232	7	0.102	7	0.036	13
KaMiKla⋆	0.078	11	0.059	15	-0.01	10	0.186	7
xxxyyyxxx	0.074	12	0.094	14	-0.013	11	0.14	8
UoR-NCL⋆	0.056	13	0.122	13	-0.043	13	0.089	12
RUG-3	0.046	14	0.137	12	–	–	–	–
suzuki	0.042	15	0.14	11	-0.018	12	0.003	15
akkhan1871	0.008	16	-0.006	16	-0.06	14	0.09	11
MaChAmp	-0.164	17	-0.131	17	-0.195	16	-0.167	16

Table 6: Sub-task 2 results for each team/user ordered by overall ρ along with results for each language. Team/user names marked with ⋆ have submitted their system description.

they first normalized the scores to be between zero and one, then performed an inverse transformation to get the final output. In addition, for this second task, they trained 10 models per language (each with its one training split) and used the median result as their final prediction.

Zhou et al. (2022) (SPDB) participated only in sub-task1, using a different ensemble system for each language. For Italian and French, the system combines the results of 10 ERNIE-M models (Ouyang et al., 2021) obtained by applying a cross-validation process; for English, the authors combined the predictions made by ERNIE-M and DeBERTa-v3 (He et al., 2021). They also enlarged the distributed PreTENS training set using *i)* two different translators (Google and Baidu) to translate the English sentences into French and Italian, thus increasing the diversity of data, and *ii)* the English and French version of the XNLI dataset (Conneau et al., 2018).

van den Berg et al. (2022) (RUG-1-pegasussers) participated only in sub-task1 using English BERT base (Devlin et al., 2019) which they fine-tuned, experimenting with multiple approaches to expand the training data. In particular, they augmented the data by adding new English sentences that contained new category templates, new words instantiating the templates, new words previously used exclusively as hyponyms, inversions of the arguments involved in the taxonomical relation, paraphrases automatically generated. The acceptability labels of Italian and French sentences were predicted by translating the sentences into English, in order to process them with the English BERT.

Markchom et al. (2022) (UoR-NCL), for both sub-tasks, experimented with fine-tuning different monolingual versions of the BERT-based model (Devlin et al., 2019), i.e. DistilBERT-Base-Uncased for English (Sanh et al., 2019), BERT-Base-Italian-XXL-Uncased for Italian, FlauBERT-Base-Uncased for French. The authors relied on the distributed training data to fine-tune the models using specific loss functions, binary cross-entropy loss for sub-task1 and mean squared error loss for sub-task2.

5 Results and Discussion

Almost all teams submitted their runs for the three languages considered. Tables 5 and 6 show, for each sub-task, the top submissions received from each team, along with the baseline scores. Team names marked with ⋆ represent teams that have submitted system description papers. The *Rank* column reports the position of the team in the ranking for global and language-specific scores.

Task 1 evaluation: To better understand the performances per construction of the models submitted, we report the average F1-macro ($\pm std$) of the top-3 submissions per language, in Table 7.

Our results show that English — the most resource-rich language in terms of computational models and data — outperforms the Italian and French models for correctly predicting presuppositions in these constructions. However, the English model performed below French in the UNLIKE construction (e.g. "Unlike trees, {*oaks /

CONSTRUCTION	EN	FR	IT	Avg_lang
DRATHER	**94.9** (±0.04)	90.9 (±0.03)	89.3 (±0.03)	91.70
COMPARATIVES	**94.2** (±0.05)	87.4 (±0.03)	88.2 (±0.02)	89.93
EXCEPT	**88.8** (±0.09)	88.3 (±0.05)	84.1 (±0.05)	87.07
UNLIKE	87.4 (±0.07)	**88.4** (±0.05)	84.6 (±0.05)	86.80
BUTNOT	**89.3** (±0.07)	78.0 (±0.15)	81.5 (±0.03)	82.93
PREFER	**86.5** (±0.1)	83.5 (±0.05)	78.1 (±0.01)	82.70
ANDTOO	**84.4** (±0.13)	74.6 (±0.14)	77.5 (±0.0)	78.83
PARTICULAR	**94.3** (±0.04)	45.3 (±0.0)	86.7 (±0.04)	75.43
TYPE	**75.8** (±0.14)	66.8 (±0.08)	72.2 (±0.12)	71.60
GENERALLY	45.5 (±0.0)	**75.2** (±0.12)	46.7 (±0.01)	55.80

Table 7: Average macro F-measure of the top 3 participants per construction in sub-task 1 (binary classification). The standard deviation between the top 3 submission are in (.). Best results per construction are in bold.

animals} are often mentioned in this text", presupp. A1⊭A2), and does quite poorly in GENERALLY ("I like oaks, and more generally {trees / *animals}", presupp. A1<A2). Note that neither constructions were present in the training set. Italian is aligned with English, while interestingly the French model seems to be capable of more accurate generalizations in both cases.

Task 2 evaluation: To gain a better understanding of the models generalization abilities in this task, we computed the Root Mean Squared Error (RMSE) between the gold value and the average predicted value by the first three teams classified. This data is shown in Figure 2 for each construction and for each language, along with the average value across languages. As it can be seen, the resulting picture contrasts substantially with that of sub-task1. The TYPE and PARTICULAR constructions, among the worst in the first sub-task, have the lowest error in sub-task2. The second task sees a substantial drop of ANDTOO (now the worst case) and COMPARATIVES (one of the best constructions in sub-task1).

We also observe distinctions between languages across the two tasks. In particular, while the average performance for English across constructions is the highest in sub-task1, the French models obtained on average the best results in sub-task2. The success at predicting the presuppositional knowledge triggered by the same construction changes in the two sub-tasks (which are often demanded to different models in the various teams). For example, the French models are the best at classifying the acceptability label for the UNLIKE construction but are the worst in predicting the human score for the same construction. Conversely, COMPARATIVES turn out to be among the easiest constructions for the English models in sub-task1 but are the most mispredicted English type in sub-task2.

Quite interestingly, the presence of a construction in training doesn't always guarantee better performances. Two notable examples are represented by DRATHER and PARTICULAR. Despite being absent in the training set of the corresponding tasks, the first obtains the highest F1 score (for English and on average) and the second is among the top-predicted constructions in sub-task2. We leave a more thorough analysis of the systematicity of these trends to future work, where we will also consider the linguistic variants for each construction and the semantic categories of the nominal arguments involved.

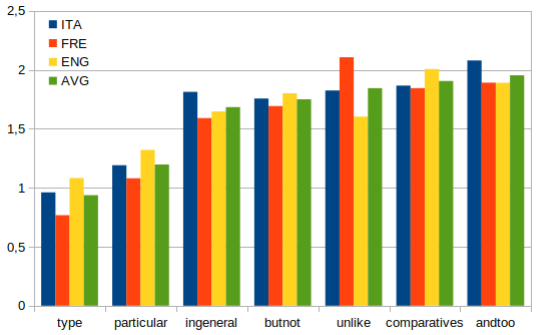

Figure 2: sub-task2: Root Mean Squared Error (RMSE) averaged across the top 3 participants per construction (lower is better). Constructions are ordered per average RMSE values across languages.

6 Conclusion

The SemEval2022 task3 – PreTENS offered two sub-tasks aiming to investigate the effectiveness of computational models to detect a certain type of presuppositional failures induced by specific constructions. The task attracted a total of 21 teams, from both academia and industry. The findings showcases the power and ubiquity of large self-supervised pre-trained models in mono- or multi-

lingual settings. Despite this apparent uniformity, the participants chose to use and combine the models in very different and creative ways, giving rise to a range of scores from 70.54% to 94.49%.

The outcomes of the task highlights the ability of these transformer models to generalize to new/unknown construction in the test sets, but also the presence of intriguing differences in specific constructions and languages (e.g. in the binary task *A1 and more generally A2* reaches a 75.2 F-measure in French but a 46.7 in Italian). Also worth further investigation is the lower correlation between the binary judgments and the human ratings in English — probably reflected in the .05 drop seen in the sub-task2 global score for this language, compared to French.

The success of the task indicates a growing interest towards research on prediction models that can incorporate world knowledge and common sense, along with an understanding of the linguistic properties that condition the outcomes. We hope that this trend will continue and the PreTENS data will help researchers to probe future models for this ability. With this spirit, we make the dataset public.

References

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2022. CSECU-DSG at SemEval-2022 Task 3: Investigating the taxonomic relationship between two arguments using fusion of multilingual transformer models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Frank van den Berg, Gijs Danoe, Esther Ploeger, Wessel Poelman, Lukas Edman, and Tommaso Caselli. 2022. RUG-1-pegasussers at SemEval-2022 Task 3: Data generation methods to improve recognizing appropriate taxonomic word relations. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation (*SEMEVAL-2016*)*, pages 1081–1091.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of International Conference on Learning Representations*, page OpenReview.ne.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, volume 1, Minneapolis, Minnesota. Association for Computational Linguistics.

K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL, HLT 2018 (16th Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies)*, pages 1195–1205, East Stroudsburg, PA. ACL.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. arXiv:2111.09543. Version 2.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium.

Yinglu Li, Min Zhang, Xiaosong Qiao, Minghan Wang, Hao Yang, Shimin Tao, and Ying Qin. 2022. HW-TSC at SemEval-2022 Task 3: A unified approach

fine-tuned on multilingual pretrained model for PreTENS. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. arXiv: 2004.10827.

Thanet Markchom, Huizhi Liang, and Jiaoyan Chen. 2022. UoR-NCL at SemEval-2022 Task 3: Fine-tuning the BERT-Based models for validating taxonomic relations. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

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.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. 2019. AlBERTo: Italian BERT language understanding model for nlp challenging tasks based on tweets. In *roceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*, pages volume 2481, CEUR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Injy Sarhan, Pablo Mosteiro, and Marco Spruit. 2022. UU-TAX at SemEval-2022 Task 3: Improving the generalizability of language models for taxonomy classification through data augmentation. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Shane Storks, Qiaozi Gao, and J. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv: Computation and Language*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

Karl Vetter, Miriam Segiet, and Klara Lennermann. 2022. KaMiKla at SemEval-2022 Task 3: AlBERTo, BERT, and CamemBERT—Be(r)tween taxonomy detection and prediction. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. ArXiv preprint arXiv:2007.00236.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. Blimp: A benchmark of linguistic minimal pairs for english. *CoRR*, abs/1912.00582.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels. ACL.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Bin Sun, Shutao Li, Kang Liu, and Jun Zhao. 2022. LingJing at SemEval-2022 Task 3: Applying DeBERTa to lexical-level presupposed relation taxonomy with knowledge transfer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yue Zhou, Bowei Wei, Jianyu Liu, and Yang Yang. 2022. SPDB Innovation Lab at SemEval-2022 Task 3: Recognize appropriate taxonomic relations between two

nominal arguments with ERNIE-M model. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.