Fine-grained Entailment: Resources for Greek NLI and Precise Entailment

Eirini Amanaki[*], Jean-Philippe Bernardy[◇], Stergios Chatzikyriakidis[*◇], Robin Cooper[◇], Simon Dobnik[◇], Aram Karimi[◇], Adam Ek[◇], Eirini Chrysovalantou Giannikouri[*], Vasiliki Katsouli[*], Ilias Kolokousis[*], Eirini Chrysovalantou Mamatzaki[*], Dimitrios Papadakis[*], Olga Petrova[*], Erofili Psaltaki[*], Effrosyni Skoulataki[*], Charikleia Soupiona[*], Christina Stefanidou[*]

[*]Department of Philology, University of Crete

{philp0898}@philology.uoc.gr, {stergios.chatzikyriakidis}@uoc.gr

{philp0899,philp0929,phil5816,philp0900,phil5647,philp0928}@philology.uoc.gr

{philp0883,philp0916,philp0861,philp0862}@philology.uoc.gr

[◇]Centre for Linguistic Theory and Studies in Probability, FLoV, University of Gothenburg

{name.surname}@gu.se

Abstract

In this paper, we present a number of fine-grained resources for Natural Language Inference (NLI). In particular, we present a number of resources and validation methods for Greek NLI and a resource for precise NLI. First, we extend the Greek version of the FraCaS test suite to include examples where the inference is directly linked to the syntactic/morphological properties of Greek. The new resource contains an additional 428 examples, making it in total a dataset of 774 examples. Expert annotators have been used in order to create the additional resource, while extensive validation of the original Greek version of the FraCaS by non-expert and expert subjects is performed. Next, we continue the work initiated by (Bernardy and Chatzikyriakidis, 2020), according to which a subset of the RTE problems have been labeled for missing hypotheses and we present a dataset an order of magnitude larger, annotating the whole SuperGlUE/RTE dataset with missing hypotheses. Lastly, we provide a de-dropped version of the Greek XNLI dataset, where the pronouns that are missing due to the pro-drop nature of the language are inserted. We then run some models to see the effect of that insertion and report the results.

Keywords: Natural Language Inference, Textual Entailment, FraCaS, RTE, XNLI

1. Introduction

Natural Language Inference (NLI, or Textual Entailment, TE) has been a core task in Computational Semantics from its early symbolic years, all the way to the present Deep Learning (DL) era. Indeed, the centrality and importance of NLI has been acknowledged early on by Cooper et al., arguing that NLI is the crux of Computational Semantics, aptly stating that "inferential ability is not only a central manifestation of semantic competence but is in fact centrally constitutive of it" (Cooper et al., 1996). This acknowledgement of the centrality of NLI has continued up to now, with NLI being one of the core tasks for Natural Language Understanding (NLU) and central to NLU benchmarks like GLUE (Vendrov et al., 2016) and SuperGLUE (Wang et al., 2019). To give a further example, one of the most cited papers in NLI (Bowman et al., 2015), argues that understanding inference about entailment and contradiction, in effect the task of NLI, is an important aspect for constructing semantic representations, while on a more practical note, Nie et al. (2020a) note that NLI is arguably the most canonical task in NLU.

Despite the great success in tackling the task of NLI in recent years, questions have started to develop about the efficiency of existing NLI datasets to train good models for NLI. For example, Chatzikyriakidis et al. (2017) have argued that the community should strive for datasets representing data from multiple domains and further include more instances of inference. This plea, as (Poliak, 2020) correctly notes, has been taken into consideration by the community, and indeed a lot of effort has been put in creating more diverse datasets in the last years. Another issue that has arisen w.r.t. dataset development is annotation artifacts, i.e. datasets that contain artifacts due to the way they are constructed, that are leveraged by the models in order to obtain good accuracy. In effect, the models are using low-level heuristics that should not play a role in solving the task. For example, (Poliak et al., 2018) have shown that artefacts and statistical irregularities can help the models perform well on the NLI task, even when only trained on the hypotheses (hypothesis-only). A lot of similar research has verified this: Pham et al. (2020) show that NLI models are not sensitive to word-order, nor to datasets corruption by random POS (part-of-speech) drop (Talman et al., 2021). In contrast, some models seem to be sensitive to changes that should not affect their performance. For example, Glockner et al. (2018) show that the replacement of words with mutually exclusive hyponyms or antonyms hurts performance, while Talman and Chatzikyriakidis (2019) show that models do not generalize well when trained and tested on different NLI datasets.

In this context, the community has tried to come up with responses to these challenges. In terms of dataset

creation, a body of research has been arguing for more diverse resources for NLI, as well as the need for datasets that are clean from annotation artefacts. As regards the former, this led to the development of more fine-grained datasets. For example, datasets that test for implicature and presupposition (Jeretic et al., 2020), Numerical/Quantifier reasoning (Kim et al., 2019; Richardson et al., 2019), Monotonicity Reasoning (Yanaka et al., 2019; Richardson et al., 2019), Comparatives, among many others.[1] As regards the latter, work on using adversarial techniques in dataset creation has led in the development of datasets much less prone to annotation artefacts. The Adversarial NLI dataset (Nie et al., 2020b) is an example of such a dataset.

One of the things directly connected to creating diverse NLI datasets, concerns multilingual NLI platforms. There is, of course, the XNLI dataset (Conneau et al., 2018), and also a number of other attempts to produce multilingual datasets for NLI for various languages (Hu et al., 2020; Wijnholds and Moortgat, 2021; Magnini et al., 2014), but in general most of the existing datasets are only in English.

In this paper, we offer a number of fine-grained resources for NLI, two for multilingual NLI, in specific for Greek, and one for precise entailment. More precisely, the paper will report the following work:[2]

- An extension of the Greek version of the FraCaS test suite that includes semantic inferences that are based on idiosyncratic features of Greek syntax. The extension makes the dataset double the size of the original.
- Validation of the original FraCaS test suite for Greek using experts and non-experts against the original annotations and result reporting.
- Completing the work in Bernardy and Chatzikyriakidis (2020) by providing the missing hypotheses (when they exist) for the SuperGLUE RTE dataset. Missing hypotheses refer to information needed to draw an inference, e.g. background knowledge, real-world knowledge, that is however missing in the premises.
- Create a version of the Greek XNLI dataset where all dropped pronouns are inserted, in effect a de-pro-dropped version of Greek. We do this in order to check whether performance of NLI models for Greek is affected if we do so, given that pre-trained language models are trained on English and are subsequently fine-tuned.

2. Methods

2.1. Extending the Greek FraCaS

The first part of the project involves the extension of the original Greek FraCaS test suite for Greek. What we wanted to achieve is an additional set of inference cases that are dependent on the syntax of Greek. Given that these cases are not so easily found in real-world data, we decided to first use expert constructed changes, focussing on the range of pattern variation for this study. The original Greek FraCaS is a translation of the English FraCaS and has been developed as part of the multi-fracas project at the University of Gothenburg.[3] The additional inference cases added include language dependent syntactic constructions that most of the time do not appear in translations of semantically similar inference cases from English to Greek. To give an example, the additive use of the coordinator *ke* rarely appears in translations of the focus associating operator "too" or "also", but rather appear with the insertion of the element meaning "also" "episis". Other cases include modal discourse markers expressing doubt like "taha" and inferences involving clitic clusters. In more detail, the extra categories added to the suite are as follows:[4]

1. Coordinator ke
    - This involves different uses of the *ke* "and" coordinator in Greek: normal conjunction, both interpretation and additive interpretation among others.
2. Negative Polarity Items
    - Inferences involving a number of negative polarity items in Greek. These include: the semantic negative operator *den* or *min* "not" followed by the NPIs *puthena* "nowhere", *kanenas* "nobody", *tipota* "nothing", *den* followed by *pote* "never" or *kan* "even", and *den* followed by *oute kan* "not even" or *oute* "neither" in embedded sentences. Also, NPIs without a negative operator: *oute kan*, *oudeis* "no one", *kanena* "anything" (existential), and *pouthena - tipota* "nowhere - nothing". There is a section with minimizers, free choice items and PPIs: *mia stalia* "a little", *kati* "something", *opoiondipote* "whoever", *toulaxiston* "at least" and *mono* "just" (Giannakidou (2011)). Lastly, there are inferences with NPIs that mean in dialogues, which highlight idiosyncrasies of Greek because include possible premises of natural speakers such as: *oute kan, pouthena kai tipota* "nothing and nowhere", and *thelondas kai min* "wanting or not".
3. Polydefinites

[1] See (Poliak, 2020) for a complete survey on NLI datasets.

[2] All resources can be found at: `https://github.com/GU-CLASP/LREC_2022/tree/main/datasets`.

[3] .

[4] Note that the list of idiosyncratic constructions that are covered in the test suite is not exhaustive. Such an exhaustive list needs further work in order first to decide which these constructions are, followed by creating examples of inference that they are involved.

- Cases that a noun is modified with an adjective and before each phrase the definite article is added (Kolliakou, 2004). While polydefinites can have a variety of semantic uses, we chose only those that have an upward entailment, because those have the most clear-cut reading among speakers.

4. Discourse Markers

   - Inferences involving three different discourse markers in Greek. The discourse markers used are the following: *siga*, *taha* and *ke kala*. *Sigs* is an adverb literally meaning "slowly", but in Greek it is used to express doubt meaning "it is doubtful" and it is associated with negation (Onufrieva, 2019). The word *taha* is an adverb meaning 'supposedly' as does the phrase *ke kala* which literally means "and well".

5. Clitics

   - This involves examples where the inference depends on weak object pronouns, for example cases of clitic clusters, where changing the case marking of the weak object pronoun gives rise to different inference patterns, e.g. the difference between an argumental and an ethical dative interpretation (*mu/me magirepse* "s/he cooked for me/ s/he cooked me").

2.2. Validating the FraCaS

The second part of the project involves the validation of the original FraCaS test suite against crowds of experts and non-experts. The validation was performed as a controlled crowd-sourcing data collection task using the Semant-o-matic tool[5] which is used for collection of semantic judgements both by targeting particular groups of participants through advertising experiment locally or on social media (as in traditional experiments and annotation tasks) or reaching out to a larger pool of participants using Amazon Mechanical Turk (Dobnik and Åstbom, 2017; Rajestari et al., 2021). In addition to the task data, questions about the participant background can also be included.

In the current data collection task all examples of the original FraCaS (346) were used. Each was presented as one of more statements (representing premises) and a question corresponding to the conclusion. Participants were instructed to answer the question by only considering information presented in the statements (the purpose was to limit the effect of background knowledge) by choosing one of the three possible answers: "Yes", "No" and "Don't know". The presentation of FraCas examples was randomised for each participant. Each participant was given a chance to provide answers to all 346 examples but there was no requirement to answer all of them as they were allowed to break the task at any time. Note that one can translate this result into a probabilistic version of the FraCaS, if they wished so: the categorical judgements over a set of participants can be translated straightforwardly to probability: the frequency by which annotators make a particular choice is the likelihood that an average annotator would make that choice.

The data was collected from subjects connected with the University of Crete in December 2021 where 175 participants were recruited among students and their social connections. Participants were asked whether they have studied linguistics before. If they answered "yes" they are considered experts (86, 49.14%) and non-experts otherwise (89, 50.86%). In total, they have provided 7,576 judgements which on average makes 21.9 judgements per FraCas example. Experts provided 3,145 judgements (41.51%) while non-experts provided 4,431 judgements (58.49%).

2.3. Precise RTE 2.0

The third part of the project involves the continuation of the work by Bernardy and Chatzikyriakidis (2020). There the authors attempt to give a precise platform for textual entailment, by taking a fraction of the RTE platform and annotate them with missing hypotheses.

We have selected all problems from the SuperGLUE/RTE task corpus which were marked as "YES" (i.e. entailment holds). The problems were not further selected nor doctored by us. The problems were then re-rated by masters students in linguistics (in Bernardy and Chatzikyriakidis (2020) experts in linguistics and logic were recruited). For most problems, three subjects were consulted (13 problems were rated by 4 subjects). More precisely, the experts were instructed to reconsider each problem and be especially wary of missing hypotheses, i.e. information used in order to carry out an inference that is however missing in the text. If they considered the entailment to hold, we gave the instruction to optionally mention any additional implicit hypothesis that they would be using. Similarly, if they considered that there was no entailment in the problem, they were suggested to (optionally) give an argument for their judgment — thereby also indirectly indicating missing hypotheses.

2.4. De-dropped XNLI

In the fourth part of the project, we investigate the effect of pro-drop in the performance of NLI models. For this reason we developed the augmented dataset depro-XNLI, where all the Greek examples have been changed by inserting all the pronouns that are missing, given the pro-drop nature of the language. We took the English cases as the basis, and inserted all pronouns that are present in English, but not in the Greek translation (see Table 1). A note on terminology here: we will be using the words de-drop/de-dropped for the pro-

[5] http://www.dobnik.net/simon/semant-o-matic/

cess/result of making a pro-drop language non pro-drop by inserting the missing pronouns.

	Premise	Hypothesis
English	*I think that's why I remember that.*	*I didn't remember it at all.*
Greek	*Νομίζω αυτός είναι ο λόγος που το θυμάμαι αυτό*	*Δεν το θυμήθηκα καθόλου*
Greek de-drop	**ΕΓΩ** νομίζω αυτός είναι ο λόγος που το θυμάμαι αυτό	***ΕΓΩ*** *δεν το θυ μήθηκα καθόλου*

Table 1: First row: Original English pairs. Second row: Translation to Greek as found in XNLI. Third row: pronoun insertion

3. Results and Analyses

3.1. Extended Greek FraCaS (EX-GR-FraCaS)

The new extended FraCaS dataset for Greek includes 774 examples of inference and can be seen as including two main parts: the existing original part[6], which is the translation of the original English FraCaS test suite into Greek and the second part, our addition, which includes a total of 428 further examples of inference that involve idiosyncratic features of Greek syntax according to the categories as these are specified in 2.1..[7] Furthermore, the original FraCaS test suite is highly imbalanced between the three categories. One can clearly see that from 3.1., where there is a clear dominance of YES examples, which take more than half the suite, approximately 0.27% are NO examples, and UNK examples are very few, comprising approximately 0.09% of the suite. Note that the original FraCaS has an additional category created by MacCartney and Manning (2007) in order to deal with defective examples that were either missing the hypothesis, or examples that had non-standard answers (e.g. Yes, on one reading) etc. This is not a negligible part of the suite as it comprises approximately 12% of the suite. The extension of the dataset is much more balanced w.r.t the three inference categories, with the YES examples comprising approximately 35% of the dataset, NO examples approximately 31%, and UNK examples approximately 34%. There are no undefined examples. The results are shown in 3.1.. Three examples from the new dataset are shown below. One involves *kanenas* "nobody", the other one *taxa* "supposedly" and the last one has to do with *kai* "and":

(1) A Yes example from the EX-GR-FraCaS test suite.

P1 Δεν ήρθε κανένας στη σημερινή παράσταση.
Nobody came at today's performance.

P2 Μόνο ο Γιώργος.
Just Giorgos.

Q. Ήρθε ο Γιώργος στη σημερινή παράσταση;
Did Giorgos come the today's performance?

H. Ο Γιώργος ήρθε στη σημερινή παράσταση.
Giorgos came at the today's performance.

Label Ναι.
Yes.

(2) An No example from the EX-GR-FraCaS test suite.

P Κοιτούσε συνέχεια το κινητό του, δεν τηλεφώνησε καν η μαμά του.
He kept looking at his phone, even his mom didn't call.

Q. Είναι αληθές, ότι η μαμά δεν τον καλεί συνήθως;
Is it true, that mom does not usually call him?

H. Η μαμά δεν τον καλεί συνήθως
Mom does not usually call him.

Label Όχι.
No.

(3) An UNK example from the EX-GR-FraCaS test suite.

P Ο Γιώργος τάχα μου τους έβλεπε πρώτη φορά στη ζωή του.
Giorgos supposedly saw them for the first time.

Q. Ο Γιώργος τους έβλεπε πρώτη φορά στη ζωή του;
Did Giorgos see them for the first time ever?

H. Ο Γιώργος τους έβλεπε πρώτη φορά στη ζωή του.
Giorgos saw them for the first time ever.

Label Δεν ξέρω.
I don't know.

3.2. Validation of the FraCaS

Figure 1 shows the results of the FraCaS validation by human judges (see Section 2.2.). The aim of the eval-

[6] https://github.com/GU-CLASP/multifracas/blob/master/fracas_greek_final_ipa_team_crete.xml.

[7] https://gu-clasp.github.io/multifracas/fracas_greek_extended_team_crete.xml

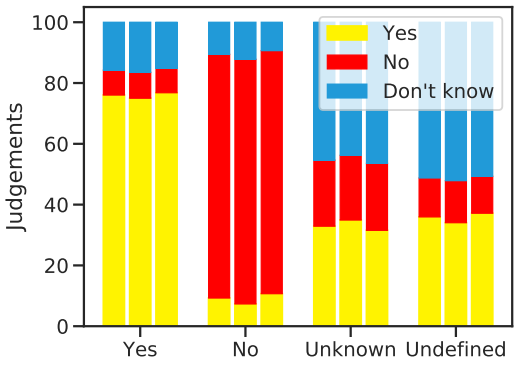
	FraCaS (original)	Addendum	EFraCaS
E	180	153	333
C	94	130	224
UNK	31	145	176
UND	41	0	41
TOTAL	346	428	774

Table 2: E stands for Entailment problems, C for Contradiction problems, UNK for neutral problems and UND for undefined. The Addendum are the extra examples added to the original Greek FraCaS, and EFraCaS the concatenation of the original Greek FraCaS and the Addendum.

uation is to examine distribution of judgments for different FraCaS categories and whether the distributions are affected by the bias from being familiar with the task. Natural language examples allow different interpretation of premises and conclusions leading to different judgments of inference, for example due to lexical ambiguity of words. This is most clearly expressed in the category "undefined". There may also be a difference in the way experts and non-experts understand inference in natural language. The horizontal axis shows the answer provided in the dataset by their designers and the vertical axis shows a percentage bar of the answers provided by human judges. For each FraCaS label we provide three bars which represent (i) all answers, (ii) expert answers, and (iii) non-expert answers. Note again that the original FraCas is imbalanced in the distribution of ground-truth labels. Out of 346 examples, there 203 (58.67%) "yes" answers, 33 (9.54%) "no" answers, 98 (28.32%) "unknown" answers and 12 (3.47%) "undef" answers. The undefined answers are difficult cases for which it was not possible to assign a different label unambiguously.

Overall there is a strong agreement with the FraCas score on "yes" and "no" classes. Sometimes examples of the yes and no classes are labelled as "unknown" and "no", possibly because participants might be bringing in additional background knowledge to resolve inference. The reason for this might be lexical or structural ambiguity of individual examples. For the examples labelled as "unknown" there is a participant bias to provide either a "yes" or "no" answer. Interestingly, this bias is lower with the "undef" label, thus those those cases that allow alternative interpretations.

A comparison of answers provided by participants who self-reported to have studied linguistics (second column) versus those who have not (third column) reveals that there are no differences between them. A χ^2 test finds no significant difference between "yes" ($p = 0.3791$), "no" ($p = .1508$), "unknown" ($p = 0.2573$) and "undef" ($p = 0.8590$) answers of linguists and non-linguists. This indicates that prior linguistic training does not have a bias on the performance on this general inference task for which no linguistic training is required. Note that the status of linguistic expertise is self reported and that participants answering this question with "yes" might have had different backgrounds and degrees of linguistic training.



(a) all judgements

Figure 1: Results of the FraCaS validation through crowd-sourcing. Each FraCaS label on the horizontal axis is associated with three bars which represent (i) all answers, (ii) expert answers, and (iii) non-expert answers.

3.3. Precise RTE 2.0

In the process, we have gathered a total of 3760 judgments, 593 missing hypotheses and 331 explanations for negative judgments. The entailment judgments are found in Fig. 2.

Despite all original problems being classified as "yes" by the creators of the RTE test suite — we find here that on average, one subject in 5 is likely to cast a doubt over this "yes". Here, we count as a doubt either a response of "no" or "yes" with missing hypotheses.

"Yes if ..." vs "No because ..."? We elected to group those categories in our summaries, because the classification between "yes" with missing hypotheses and "no" is a tenuous one. Indeed, experts often find the same missing hypotheses but classify the problems differently (as "yes" or "no").

We find that missing hyphotheses tend not to be discovered by all subjects. As evidence, the agreement factor (Fleiss' Kappa) when grouping answers in the doubtful/certain categories is κ=0.16.

Another way to look at the data is to count the number of experts casting doubt on an entailment problem. In Fig. 3, we show the distribution of number of experts casting doubt on entailment, over all problems, as a histogram.

To sum up,

1. Perfect agreement (0 or 3 doubts) occur in 47 percent of cases.

2. The probability of having a three doubts being cast is the lowest.

Type	Count	Ratio
Yes, with no missing hypothesis	2636	0.70
Yes, with missing hypotheses	593	0.16
No, with no explanation	200	0.05
No, with explanation	331	0.09
Total of doubtful entailment	734	0.20
Total of any type	3760	1

Figure 2: Number of responses by type

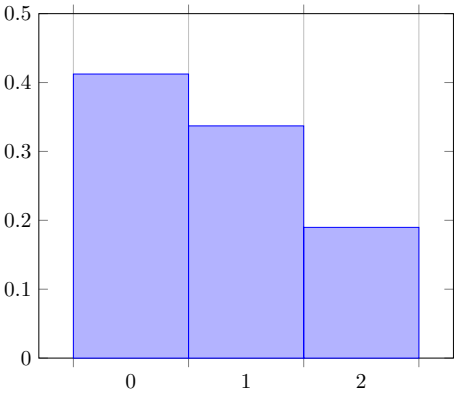


Figure 3: Distribution of the number of doubtful subjects.

We find this level of agreement indicative of a good level of reliability. Additionally, with three experts per problem, we are likely to discover most missing hypotheses and incorrect entailments.
In this setting, we have found that subjects were less likely to cast doubt on entailment than Bernardy and Chatzikyriakidis (2020). We conjecture that this is because master students are less likely to discover gaps in reasoning than the more seasoned experts (PhD or professors in linguistics or logic) consulted by Bernardy and Chatzikyriakidis (2020). The size of the sample might also have an effect, given that it is ten times the size of the original. It would be interesting to repeat the experiment with more seasoned experts or the other way around, i.e. use the smaller sample with less experienced annotators. In any case, the other issue that this discrepancy between the number of missing premises identified in Bernardy and Chatzikyriakidis (2020) and the smaller number we have found in this study shows, is that the task of finding missing premises is rather open-ended and can go to different levels of fine-grainedness. This further shows the problem with some cases of inference, namely that a lot of missing knowledge has to be recognized by the model and/or find a way to make the inference in a way that resembles this kind of reasoning under hidden premises.

3.4. De-dropped XNLI

We evaluate the effect of inserting pronouns in the Greek XNLI dataset to investigate whether the pro-drop differences of the languages have an effect in the performance of the models. Our goal here is to not to make other languages similar to English, but to investigate the importance of pro-drop in such tasks, if any. We use the XLM-RoBERTa (Conneau et al., 2019) model trained on the English MNLI dataset (Williams et al., 2017). Our model uses max-pooling over the word representations to obtain a sentence representation. We found this method more effective than taking the CLS representation. In the experiment we evaluate how effective transfer learning is when presented with unusual syntax (that does not alter the meaning) in Table 3.

Data	Accuracy
Original	75.0
De-drop	74.8

Table 3: Results on the original XNLI data and the de-dropped data.

The results show a small drop in accuracy of 0.2 percentage points. This indicates that for models trained on English NLI examples, when transferring the knowledge to Greek, models are able to account for examples where dropped pronouns have been added back to the sentence. However, as can be seen in Table 1, adding the pronouns may result in a lexical overlap between the premise and hypothesis which the model can exploit. For this reason, we also test the scenario where only the premise or the hypothesis have the inserted pronouns in Table 4.

Premise	Hypothesis	Accuracy
Original	De-drop	68.8
De-drop	Original	68.9

Table 4: Results when de-dropping either the premise or hypothesis.

When only one of either the premise or hypothesis have the pronoun inserted we see that the performance degrades by 6.2 percentage points. This indicates that while some cases of inserted pronouns are handled correctly by the model, it also changes the label on some examples. In addition to highlighting issues NLI models have with inserting pronouns, this also shows that the models also rely on the lexical overlap between the premise and hypothesis, even when the overlap is non-consequential pronouns.

4. Conclusion and Future Work

In this paper, we provided a number of resources for Greek NLI, as well as precise entailment. More specifically, we extended the FraCaS test suite for Greek to further include cases of inference that are dependent on language specific syntax. The resulting test suite is double the size of the original one. We believe that such an extension can be taken as a starting point for developing multilingual NLI datasets that cover the wealth of reasoning patterns in interaction with language dependent syntax.

Next, we performed a validation of the original FraCaS test suite for Greek against both experts and non-experts. The results show a number of good agreement with the original test suite, even though some digressions exist, especially for the UNK category. No significant difference between expert and non-expert annotation has been found.

Connected to the previous is the finding that cases of entailment in datasets like the RTE involve hidden premises that are implicitly taken into consideration in the inference process. Following the work by Bernardy and Chatzikyriakidis (2020), we provided annotation of these missing premises for the whole RTE as this is found in SuperGLUE.

Lastly, we presented a variation of the XNLI Greek dataset, where all pronouns included in the original English examples and are missing in the Greek version, due to the pro-drop nature of the language, are introduced. This leads to the creation of a de-dropped XNLI dataset for Greek. We wanted to test the hypothesis of whether this data augmentation/corruption will have an effect on model performance. No effect was found when the new de-dropped dataset was used. However, an effect was found when we used a hybrid format: a) the premises are in the original format but the hypotheses in the de-dropped form and b) vice versa. In these cases, we found a significant drop in performance which points to the system exploiting various lexical overlap cues in deciding inference.

We believe that what we have proposed in this paper can be extended to multiple languages, but also to multiple task investigations. As regards the former, we believe that the idea of providing examples of inference based on idiosyncratic syntax of the target languages is a promising way towards better multilingual NLI and we hope that more researchers will pick up on this idea. The next step is to ground these new example cases in natural data. This is what we plan to do in future work. The results in the validation task, as well as the annotation for missing inferences brings out the fact that inference is not one consistent thing, but rather varies depending on context, expertise, domain and so on. It also brings out the fact that the annotation guidelines are extremely crucial in the results one gets w.r.t inference. One promising way to further extend this work is to design systems that can automatically infer hidden premises given a premise, a hypothesis and their label. Lastly, w.r.t the last part of the paper, where a de-dropped version of the Greek XNLI dataset was presented, such a dataset or similar dataset can investigate more theoretical issues w.r.t to various linguistic features that vary between languages, pro-drop being one of them. This will eventually lead in NLP working closer with Theoretical Linguistics in order to investigate theoretical claims made w.r.t these varying features.

5. Acknowledgements

Jean-Philippe Bernardy, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, Adam Ek and Aram Karimi are supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

Bernardy, J.-P. and Chatzikyriakidis, S. (2020). Improving the precision of natural textual entailment problem datasets. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6835–6840.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Lluís Màrquez, et al., editors, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pages 632–642. The Association for Computational Linguistics.

Chatzikyriakidis, S., Cooper, R., Dobnik, S., and Larsson, S. (2017). An overview of natural language inference data collection: The way forward? In Proceedings of the Computing Natural Language Inference Workshop.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. (1996). Using the framework. Technical report.

Dobnik, S. and Åstbom, A. (2017). (Perceptual) grounding as interaction. In Volha Petukhova et al., editors, Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue, pages 17–26, Saarbrücken, Germany, August 15–17.

Giannakidou, A. (2011). Negative and positive polarity items. De Gruyter Mouton, 2:1660–1712.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. In The 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, July.

Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., and Moss, L. S. (2020). Ocnli: Original chinese natural language inference. In Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 3512–3526.

Jeretic, P., Warstadt, A., Bhooshan, S., and Williams, A. (2020). Are natural language inference models imppressive? learning implicature and presupposition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8690–8705.

Kim, N., Patel, R., Poliak, A., Xia, P., Wang, A., McCoy, T., Tenney, I., Ross, A., Linzen, T., Van Durme, B., et al. (2019). Probing what different nlp tasks teach machines about function word comprehension. In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), pages 235–249.

Kolliakou, D. (2004). Monadic definites and polydefinites: their form, meaning and use. Journal of linguistics, 40(2):263–323.

MacCartney, B. and Manning, C. D. (2007). Natural logic for textual inference. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 193–200. Association for Computational Linguistics.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Pado, S., Stern, A., and Levy, O. (2014). The excitement open platform for textual inferences. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pages 43–48.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020a). Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, et al., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4885–4901. Association for Computational Linguistics.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020b). Adversarial nli: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901.

Onufrieva, E. (2019). Συντακτικοί φρασεολογισμοί με σημασία άρνησης στη νέα ελληνική. In Μελέτες για την ελληνική γλώσσα 39, pages 1143–1158.

Pham, T. M., Bui, T., Mai, L., and Nguyen, A. (2020). Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? arXiv preprint arXiv:2012.15180.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 180–191. Association for Computational Linguistics.

Poliak, A. (2020). A survey on recognizing textual entailment as an nlp evaluation. arXiv preprint arXiv:2010.03061.

Rajestari, M., Dobnik, S., Cooper, R., and Karimi, A. (2021). Very necessary: the meaning of non-gradable modal adjectives in discourse contexts. In Peter Ljunglöf, et al., editors, Selected contributions from the Eighth Swedish Lan- guage Technology Conference (SLTC-2020), volume 184 of NEALT Proceedings Series, No. XX, Gothenburg, Sweden, 25–27 November. Northern European Association for Language Technology (NEALT), Linköping University Electronic Press: Linköping Electronic Conference Proceedings.

Richardson, K., Hu, H., Moss, L. S., and Sabharwal, A. (2019). Probing natural language inference models through semantic fragments. CoRR, abs/1909.07521.

Talman, A. and Chatzikyriakidis, S. (2019). Testing the generalization power of neural network models across NLI benchmarks. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 85–94, Florence, Italy, August. Association for Computational Linguistics.

Talman, A., Apidianaki, M., Chatzikyriakidis, S., and Tiedemann, J. (2021). NLI data sanity check: Assessing the effect of data corruption on model performance. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 276–287, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2016). Order-embeddings of images and language. In Yoshua Bengio et al., editors, 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems, pages 3266–3280.

Wijnholds, G. and Moortgat, M. (2021). Sicknl: A dataset for dutch natural language inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1474–1479.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Yanaka, H., Mineshima, K., Bekki, D., Inui, K., Sekine, S., Abzianidze, L., and Bos, J. (2019). Can neural networks understand monotonicity reasoning? In Proceedings of the 2019 ACL Workshop

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 31–40.