OYXOY: A Modern NLP Test Suite for Modern Greek

Anonymous ACL submission

Abstract

This paper serves as a foundational step towards the development of a linguistically motivated and technically relevant evaluation suite for Greek NLP. We initiate this endeavor by introducing four expert-verified evaluation tasks, specifically targeted at natural language inference, word sense disambiguation (through example comparison or sense selection) and metaphor detection. More than language-adapted replicas of existing tasks, we contribute two innovations which will resonate with the broader resource and evaluation community. Firstly, our inference dataset is the first of its kind, marking not just *one*, but rather *all* possible inference labels, accounting for possible shifts due to e.g. ambiguity or polysemy. Secondly, we demonstrate a cost-efficient method to obtain datasets for under-resourced languages. Using ChatGPT as a language-neutral parser, we transform the Dictionary of Standard Modern Greek into a structured format, from which we derive the other three tasks through simple projections. Alongside each task, we conduct experiments using currently available state of the art machinery. Our experimental baselines affirm the challenging nature of our tasks and highlight the need for expedited progress in order for the Greek NLP ecosystem to keep pace with contemporary mainstream research.

1 Introduction

It is a well known fact that the natural language processing world is running at multiple speeds. A select few languages claim the lion's share in the literature, boasting a plethora of models and a constant stream of results, while others are struggling to keep up with last year's state of the art. Meanwhile, multilingual models, despite being heralded as the end-all solution to the issue, often fall short of expectations (Wu and Dredze, 2020; Ogueji et al., 2021; Pfeiffer et al., 2021; España-Bonet and Barrón-Cedeño, 2022; Havaldar et al., 2023; Papadimitriou et al., 2023, *inter alia*). The assumption that one-size-fits-all multilingual models can effectively bridge the language gap is hard to either refute or validate, given the disproportionate distribution of training and evaluation resources among languages (Joshi et al., 2020; Yu et al., 2022; Kreutzer et al., 2022). Further muddying the waters is the dubious quality of the increasingly trending multi- and mono-lingual resources generated through minimally supervised machine translations from English (Artetxe et al., 2020; Wang and Hershcovich, 2023). While such endeavors can certainly make for good first steps, they are neither sufficient nor without risks. The wide adoption of the practice threatens resource plurality, as more and more "new" datasets are in fact old in all but language. Furthermore, it condones the accumulation of academic authority to a select few, namely the authors of the originals, promoting the unhindered perpetuation of their biases and oversights as universal across languages. Worse yet, it outsources linguistic expertise to machine labor, as we are now entrusting our automated processes with capturing the nuances of under-represented languages; exactly *those* languages that require opinionated and targeted expert attention the most.

And while a discussion on the structural causes behind the problem and the ways to incentivize

change is long overdue, here we set our aims towards something more actionable. Noting the striking absence of evaluation benchmarks for modern Greek, and the language's limited presence in multi-lingual resources, we set out to develop a linguistically motivated and technically relevant suite of evaluation tasks. This paper aims to kickstart this endeavor, while serving as an open invitation to interested parties. Concretely, we set the pace with four evaluation tasks:

1. a handcrafted dataset for inference, consisting of 1 762 sentence pairs, each pair adorned with a linguistic characterization in the form of tags *à la* SuperGlue and labeled with a subset (rather than an element) of {`Neutral`, `Entailment`, `Contradiction`}, aiming to account for all possible inference relations between premise and hypothesis
2. a structured translation of the Dictionary of Standard Modern Greek, from which we project into three tasks:
   (i) a word sense disambiguation task *à la* Words-in-Context, consisting of 117 662 phrase pairs that correspond to two usage examples for a single word, where the system is tasked with telling whether the two occurrences have the same meaning or not
   (ii) a more compact & linguistically informed version of the same task consisting of 14 416 unique phrases containing polysemous words, each word associated to a number of senses and their periphrastic definitions, where the system is tasked with telling which word sense is associated with each usage example
   (iii) a metaphor detection task, associating each of the previous phrases to a boolean label indicating whether the word in focus is used metaphorically or not

To facilitate research with these tasks, we supply accessible entry points to the raw data in the form of Python interfaces. For each task, we conduct experiments using the currently available state of the art machinery and establish baseline scores for comparisons.[1]

2 OYXOY

Inspired by Glue and SuperGlue (Wang et al., 2018, 2019), our goal is to develop a language-adapted suite that selects and extends a few key aspects of the original. Our project, which we lightly dub OYXOY (pronounced /ˈu.xu/), is not primarily focused on offering general diagnostics, but rather on highlighting the semantic, syntactic, and morphological attributes of the Greek language, and quantifying their impact on NLP systems. To that end, we present four high-level tasks that require varying degrees of lexical & sentential meaning comprehension.

2.1 Natural Language Inference

Our first task is a staple of computational semantics that has endured the test of time: natural language inference (NLI). In their most common form, NLI tasks present the system with an ordered pair of sentences (called a premise and a hypothesis), and request one of three inference relations that must hold between premise to hypothesis: `Entailment`, `Contradiction` and `Neutral`/`Unknown`. Despite its apparent simplicity and the heaps of progress in modern NLP, the conquest of NLI has proven challenging to this day. Neural systems show a tendency to abuse spurious data patterns over actually performing the (often complicated) reasoning required to solve the problem, resulting in limited generalization capacity across datasets. For our dataset, we follow Wang et al. (2018, 2019) in establishing a hierarchy of rudimentary but descriptive linguistic tags that encompass an array of phenomena that can influence the direction of inference. For a glimpse at the full hierarchy of tags used, refer to Table 2. These tags are intended to find use outside the model's input/output pipeline, providing a guide for categorizing results and drawing finer-grained quantitative evaluations. Where our dataset diverges from established practices is in providing an explicit account of inference-level ambiguities not only through the tagging but also through the labeling scheme. Rather than annotating each example pair with any *one* inference label, we instead specify *all* possible labels that may hold. To do so, we implicitly consider the product space of all possible readings of both premise and hypothesis, and construct the label set arising out of all pairwise interactions; Figure 1 shows two concrete examples under different settings.

To create the collection of samples that make up the dataset, we follow a three stage process. At the first stage, each author independently wrote a number of sentence pairs together with a sug-

[1] Data, interfaces and the code necessary to replicate our experiments is provided as supplementary material.

gested set of tags and labels. Afterwards, each author was given a collection of sentence pairs from other authors with the tags and labels hidden, and was tasked with assigning the tags and labels they deemed most appropriate. This way, we end up with four unique tag and label sets for each pair. Finally, we perform an aggregation of the proposed annotations and jointly go through any and all examples that contain at least one tag or label that does not reach a majority (i.e. counts less than three votes). We resolve disagreements by adding or removing annotations, thus ensuring internal consistency within the dataset. At the end of the process, we end up with 1 049 samples, of which 110 contain more than a single label. The dataset as a whole contains 454 `Neutral`, 414 `Entailment` and 292 `Contradiction` assignments.

In parallel to the above, we re-annotate the Greek version of FraCaS (Amanaki et al., 2022) according to our format specifications, skipping directly to the third stage of the pipeline described earlier. The derived dataset contains an additional 713 examples, revealing 30 of them as multi-labeled, with a label distribution of 264 `Neutral`, 345 `Entailment` and 134 `Contradiction`. We serve the two datasets independently, but as a single resource.

2.2 Repurposing the Lexicon

Transitioning to our next objective, a resource targeting lexical semantics, we immediately run into a roadblock. The construction of a sufficiently large dataset centered on the *word* requires a prohibitive investment of time and effort. Facing the very same challenge, contemporary contributions have established the practice of turning to either machine translation or crowd-sourced labor, with hired workers being overlooked by applied practitioners (at best, if at all). Albeit pragmatic, this approach compromises the quality of the generated resources, dismissing domain expertise in the pursuit of improved cost efficiency (a prerequisite, in turn, for quantity). As an alternative, we redirect our focus towards a frequently-overlooked traditional resource: the *lexicon*. Reputable lexica offer a rare mixture of linguistic rigor and extensive coverage virtually for free, making them a prime candidate for adaptation and repurposing into modern applications. In what follows, we showcase how this insight can be put into practice, enacting a sensible and effective way forward for under-resourced languages.

We begin by procuring a copy of the Dictionary of Standard Modern Greek (Triantafyllides, 1998).[2] The dictionary is provided in the form of a minimally structured SQL database, associating each lemma with its lexical entry, a raw text field containing a periphrastic definition and a few usage examples for each of its senses. Unfortunately, senses and examples are not structurally differentiated by the database, but are rather presented in the same field, further intertwined with supplementary details such as usage conditions, morphological information, etc. Instead, the database relies on a combination of formatting strategies, including enumeration and styling, to differentiate between definitions and examples. However, these strategies are not consistently applied across the lexicon. To make matters worse, definitions and examples are often woven together (that is, they materialize as non-contiguous strings), and can at times follow ad-hoc hierarchical arrangements. Consequently, even though the textual content effectively conveys information visually, parsing this content with traditional methods proves nigh impossible. As a workaround, and considering that parsing unstructured data is a staple task for large language models, we employ ChatGPT (Brown et al., 2020) for the problem at hand.

Our pipeline is as follows. We first utilize the existing database fields to filter the lexical entries that seem to contain at least one example. This results in a collection of 28 831 unique lemmata, each mapped to its lexical entry. We randomly sample 100 of them, which we then manually convert into a succinct and minimally structured JSON format, specifying (i) the *lemma* and (ii) a list of *senses*, each sense structured as a *definition* and a list of *examples*. We put extra effort into disentangling hierarchical senses, repeating the elided parts of non-contiguous definitions and examples and removing enumeration identifiers. The yield of this process then serves as the training set for a quick one-shot tuning of ChatGPT[3], the input being the raw text (stripped of HTML tags for token economy) and the target being the structured JSON representation. We pass all remaining entries through the trained model. From the model output, we filter out senses that contain no examples and entries that contain less than two senses, and end up with 16 079 examples spread over 7 677 senses

[2]Hosted online at `www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides`.

[3]We use model `gpt-3.5-turbo` via the fine-tuning API.

Premise	Ο Κυριάκος φίλησε την Αντιγόνη. *Kyriakos kissed Antigone.*
Hypothesis	Ο Κυριάκος και η Αντιγόνη φιλήθηκαν. *Kyriakos and Antigone kissed [each other].*
Labels	`Entailment`, `Unknown`
Tags	Lexical Entailment:Symmetry/Collectivity
Premise	Ο Γιώργος είπε στη Μαρία ότι ξέρει να παίζει κιθάρα. *Giorgos told Maria that [he/she] knows how to play the guitar.*
Hypothesis	Η Μαρία ξέρει να παίζει κιθάρα. *Maria knows how to play the guitar.*
Labels	`Entailment`, `Unknown`
Tags	Lexical Entailment:Factivity:Factive, Predicate-Argument Structure:Anaphora/Coreference

Figure 1: NLI examples 761 and 879, showcasing multiple inferences. In the first example, φιλώ [/filó/] (*to kiss*) can be a unidirectional or a reciprocal action (i.e., *to give a kiss to* vs. *to exchange kisses with*). In the second example, pro-drop allows for two possible readings, where either Giorgos or Maria can be the subject of the embedded clause.

and 2 512 entries. Finally, we manually check each and every example and entry, throwing away the occasional parsing error, homogenizing the presentation and fixing the JSON formatting as needed. The result is 14 416 examples spread over 6 896 senses and 2 326 entries, from which we derive the three evaluation tasks described in the subsections to follow.

The Role of ChatGPT Our decision to incorporate a large language model model into our data preparation process does not entail any of the epistemological risks commonly associated with generative models and/or data augmentation. In our use case, the model does not need a deep understanding of the Greek language, the expertise of a trained linguist, or the creativity required of a human annotator, as it's neither generating new examples nor annotating existing ones per se. Rather, it suffices for it to recognize the inconsistent yet intuitive hierarchical enumeration patterns present in the data, and to convert them into recurring structures with consistent formatting. Large language models' attested proficiency in this scenario align them perfectly with our needs, allowing us to utilize the authoritative resource of the lexicon while minimizing tedious human labor and cost expenditure. Indeed, our inspection of the model's output shows a generally high-quality translation, strictly faithful to the original input, with only a few minor occasional inconsistencies[4].

[4]The model is sometimes overeager, extending the output specification with additional fields, in what seems like an attempt to capture all the information provided in the raw input.

2.2.1 Words-in-Context

The first task is essentially a replica of the Words-in-Context (WiC) part of SuperGlue. It is formulated as a binary classification problem, where the system is presented with two sentences containing the same (potentially polysemous) word, and is tasked with telling whether the two occurrences correspond to the same meaning or not. In order to successfully resolve the task, the system needs a dynamic embedding strategy, capable of disambiguating words depending on their surrounding context. As such, it serves as a primitive test suite for the lexical semantic capacities of bidirectional transformers.

Obtaining the task from our dataset is trivial; it suffices to consider the sum of the product space of examples for each lexical entry (with the diagonals removed), zipped with a boolean sign indicating whether the two examples stem from the same sense. Doing so yields 117 662 data points (i.e., one order of magnitude larger than the corresponding fragment of SuperGlue), with a label ratio of 1 positive to about 6 negative.

2.2.2 Sense Selection

The above formulation is straightforward, and directly compatible with the standard sequence classification pipeline commonly employed by NLP architectures. As such, it makes for an accessible entry point for evaluation. However, it represents a dramatic simplification of the disambiguation problem, requiring two usages in juxtaposition and providing little information on *what* the sense of each usage is. Our source dataset allows us to do better. Given that we have periphrastic definitions

for all[5] the possible meanings of each word, we can reframe the task as sense selection. Given a word, the set of its possible meanings and a usage context, we can prompt a model to predict the meaning most likely employed in the given context. Using periphrastic definitions as a proxy for meaning induces a better informed and more realistic evaluation task, requiring and benefiting from high-quality contextual representations both at the lexical and the sentential level (since the word under scrutiny will now need to be contrasted to the full set of "meanings"). It is also more faithful to the source dataset, since the count of data points is now in alignment with the number of distinct usage examples (as duplication is no longer necessary). Each of the 14 416 points is associated with 3.8 candidate definitions, on average.

2.2.3 Metaphor Detection

Our projection of the raw textual entries into structured JSON entries has done away with most fields irrelevant to word disambiguation. However, we have consciously kept markers of metaphoric usage, and homogenized their presentation.[6] This enables us to filter senses (and by extension, usage examples) that are used metaphorically, providing the means for another kind of task altogether: metaphor detection. Making the simplifying assumption that metaphor is only present in those examples where the word defined is used in a metaphoric sense, we end up with 1 017 examples of metaphor (7% of the total of all examples) concentrated around 571 senses and associated with 499 entries, yielding a heavily imbalanced dataset for metaphor detection.

3 Experimental Baselines

To quantitatively evaluate the difficulty of the tasks described in the previous section, and in order to facilitate future research in this direction, we set up some experimental baselines using the current state-of-the-art machinery available for modern Greek. All our experiments rest on the tried and tested fine-tuning process for BERT-like models (Kenton and Toutanova, 2019), using Greek BERT as our universal core model (Koutsikakis et al., 2020).

[5]Excluding the ones removed by the filtering process.

[6]They are indicated with (μτφ.) in the periphrastic definition.

3.1 Natural Language Inference

Despite our efforts to create a comprehensive evaluation suite for natural language inference, the practical use of our dataset presents several challenges. First and foremost, its comparatively small size renders it unsuitable for fine-tuning purposes. This becomes especially problematic considering the lack of NLI datasets tailored specifically for Greek. Compounding these challenges is the fact that our dataset utilizes a multi-label setup, which complicates direct cross-dataset evaluations. To address these challenges, we have chosen to leverage XNLI (Conneau et al., 2018), a cross-lingual dataset for language inference of substantial size; while XNLI was not initially designed for training purposes, it presents a viable solution considering the constraints we face. We employ an iterative approaching when splitting our dataset, aiming for a 30/70 division and taking care to keep the ratio consistent for each of the linguistic tags used. We then fine-tune BERT, training on the joined test set of XNLI and the smaller of the two splits, evaluating on the dev set of XNLI, and testing on the larger split. This setup accounts for domain adaptation, while allowing us to frame the problem as multi-label classification (where the XNLI problems are "coincidentally" single-label).

Concretely, we independently contextualize the premise and hypothesis sentences, concatenate their `[CLS]` tokens and project them into three independent logits via an intermediate feed-forward layer of dimensionality 64, gated by the GELU activation function (Hendrycks and Gimpel, 2016). We train using AdamW (Loshchilov and Hutter, 2018) with a batch size of 32 and a learning rate of 10^{-5}. Despite heavy regularization (weight decay of 0.1, dropout of 0.33 and early stopping), the model is quick to overfit the training set, with development set performance lagging significantly behind (despite the matching domain). Since accuracy is no longer a suitable performance metric, owing to the multi-label setup we have adopted, we report per-class precision, recall and F1 scores over the test set instead, averaged over three repetitions. The results, presented in Table 1, are largely underwhelming, indicative of the difficulty of the dataset and confirming the inadequacy of (the Greek fragment of) XNLI as a training and evaluation resource – a fact also noted by Evdaimon et al. (2023) and consistent with the comparatively low scores of Amanaki et al. (2022). To

Label	Prec.	Rec.	F1
Unkn.	0.32±4.9%	0.41±1.0%	0.35±3.7%
Ent.	0.52±2.8%	0.46±2.7%	0.48±1.1%
Contr.	0.20±0.7%	0.26±7.6%	0.23±0.6%

Table 1: Per-label test metrics for NLI.

gain a better understanding of the trained model's behavior across different linguistic phenomena, we group samples according to their linguistic tags, and measure the average Jaccard similarity coefficient between predicted and true labels (i.e., the length of the intersection over the length of the union between the two sets). As Table 2 suggests, performance is consistently low across the board. The model seems to especially struggle with recognizing the effect of embedded clauses (regardless of whether they are restrictive or not), focus associating operators, non-intersective adjectives, hypo- and hypernymy, antonymy and negation.

3.2 Sense Disambiguation

For both variants of the sense disambiguation task, we split the dataset's examples into three subsets: a 60% training set, a 20% development set, and a 20% test set. Additionally, we designate 10% of the total lexical entries as test-only, and move the associated examples from the training set to the test set. This will allow us to evaluate the model's performance separately on in- and out-of-vocabulary examples (IV and OOV, respectively), i.e. involving words that have or have not been encountered during training.

To find the relevant word within each example, we lemmatize examples using SpaCy (Honnibal et al., 2020, model `el_core_news_sm`) and identify the element within each sequence that corresponds to the source entry's lemma, falling back to the element with the minimal edit distance if no absolute match can be found. Following tokenization, this permits us to create a boolean mask for each example, selecting only these tokens that are associated with the word/lemma of interest.

Words-in-Context For the WiC variant, we gather minibatches consisting of all examples that belong to the same lexical entry. We contextualize examples independently, and extract the representations of the words of interest by mean pooling the last layer representations of the tokens selected by each example's mask. We then compute pairwise similarity scores between pairs in the cartesian

Tag	Jaccard Index (ave.)
Logic	
Disjunction	0.32±3.2%
Conjunction	0.41±1.6%
Negation	
Single	0.30±1.6%
Multiple	0.46±5.6%
Negative Concord	0.32±0.4%
Comparatives	0.42±3.5%
Quantification	
Existential	0.43±1.0%
Universal	0.36±1.3%
Non-Standard	0.37±2.8%
Temporal	0.32±1.1%
Conditionals	0.32±3.2%
Lexical Entailment	
Redundancy	0.33±1.1%
Factivity	
Factive	0.41±2.2%
Non-Factive	0.32±4.0%
Intersectivity	
Intersective	0.38±4.2%
Non-Intersective	0.29±7.4%
Restrictivity	
Restrictive	0.28±2.9%
Non-Restrictive	0.27±4.0%
Lexical Semantics	
Synonymy	0.46±2.9%
Hyponymy	0.47±1.8%
Hypernymy	0.29±5.6%
Antonymy	0.30±3.2%
Meronymy	0.50±2.5%
Morph. Modification	0.33±1.8%
FAO	0.28±1.3%
Symmetry/Collectivity	0.44±4.1%
Predicate-Argument Structure	
Alternations	0.38±2.0%
Ambiguity	0.40±2.9%
Anaphora/Coreference	0.39±0.1%
Ellipsis	0.44±1.7%
Core Arguments	0.55±5.0%
Common Sense/Knowledge	0.36±0.3%

Table 2: Per-tag test metrics for NLI. The tag hierarchy follows along Wang et al. (2019), with few divergences. For Logic, we replace Double Negation with Multiple Negations and differentiate it from Negative Concord. We add a tag for Non-Standard Quantification, and drop the Numeral/Interval tag. For Lexical Entailment, we substitute Morphological Negation with the (more general) Morphological Modification. We subcategorize Lexical Semantics, specifying left-to-right or premise-to-hypothesis (directional) lexical relations. Finally, we merge Common Sense and World Knowledge into a single meta-tag.

product of examples by applying the dot-product operator on the extracted representations, scaling the results by the inverse of the square root of the model's dimensionality. These similarity scores serve as logits for binary cross entropy training, predicting whether the two occurrences of the word share the same sense between the two examples.

Sense Selection For the sense selection variant, we create batches by (i) sampling over training examples and (ii) constructing the set union of all related (candidate) definitions, together with a binary boolean relation specifying whether an example and a definition belong to the same entry. We then independently contextualize all examples and definitions, extracting contextual word representations for each example as before, and taking each definition's `[CLS]` token representation as a proxy for the sense's meaning. We compare each word (in the context of a single example) to each meaning using the same scaled dot-product mechanism as before, masking out invalid pairs according to the example-to-definition relation mentioned earlier. We finally obtain softmax scores for each example yielding a probability distribution over candidate meanings, which serves as the model outputs for standard negative log-likelihood training.

We train on either task using AdamW with a learning rate of 10^{-5}, a weight decay of 10^{-2} and a 25% dropout applied at the dot-product indices, and perform model selection on the basis of development set accuracy; once more, development and training set performances quickly diverge after a few epochs. At this point, we note that both tasks use the same notion of sense agreement and both our models approximate it by means of the same vector operation; their difference lies in the fact that one compares a word occurrence to a word occurrence (or: an example to an example), whereas the other compares a word occurrence to a set of "meanings" (or: an example to all candidate definitions) (Hauer and Kondrak, 2022). Intuitively, it would make sense that a model that has acquired the sense selection task should be able to perform adequately on the WiC task without further training; indeed, if two word occurrences select the same meaning (i.e., maximize their similarity to the same vector), they must also be similar to one another. To test this hypothesis, we simply apply the model obtained by fine-tuning on the sense selection task, except now recasting the test set in the form of the WiC task.

We report repetition-averaged aggregates in Table 3. Performance is not astonishing, but remains well above the random baselines for both tasks (25% for sense selection and 16.7% for WiC), indicating that the core model has some capacity for learning and generalization. Sense selection may initially appear as the more challenging of the two tasks, seeing as it involves selecting one target out of multiple options. Nonetheless, the model achieves a consistently higher absolute accuracy there; evidently, comparing one example to a fixed set of senses is easier than comparing two ad-hoc usage examples. To our surprise, we find that the task transfer setup works straight out of the box, to the point where the transfer model in fact outperforms the in-domain model without as much as recalibrating the sigmoid classification threshold. One might hypothesize that this is due to the model memoizing a fixed set of senses and their representations. However, this is not entirely the case: interestingly, accuracy now improves instead of declining in the OOV fragment of the test set. We interpret this as evidencing that the sense selection formulation produces a higher quality error signal, which induces a better informed disambiguation prior during fine-tuning, allowing the (more rudimentary) WiC task to be captured without additional effort.

3.3 Metaphor Detection

The last task, metaphor detection, is also the simplest one, being essentially a case of sequence classification. We start by filtering all entries that have at least one metaphoric sense, so as to alleviate the severe class imbalance of the full dataset. From the 499 filtered entries, we reserve 5% for use as an OOV test set. We extract all examples from all entries, and assign to each example a boolean label, indicating whether the sense the example is associated with is metaphoric or not. This produces 3 015 examples (2 856 IV and 159 OOV), with a class distribution of about 1 positive to 2 negative. We proceed with training using once more a 60/20/20 split on the IV set.

We attach a feedforward classifier to the contextualized `[CLS]` token and train using binary cross entropy, optimizing with the same hyper-parameter setup as before. Our results, presented in Table 4, showcase a good ability to recognize metaphoric senses in the words trained on, and a decent gener-

	Sense Selection		Words-in-Context		
Subset	# examples	accuracy	# pairs	accuracy[1]	accuracy[2]
IV	2 494	0.63±0.20%	8 274	0.50±0.41%	0.51±1.7%
OOV	1 289	0.64±0.41%	9 954	0.48±1.77%	0.54±0.2%
Total	3 784	0.63±0.29%	18 678	0.49±1.09%	0.53±0.86%

[1] In-domain evaluation of the words-in-context model.

[2] Transfer evaluation of the sense selection model.

Table 3: Test set sizes and performance metrics for the two sense disambiguation tasks.

Subset	# Examples	**Accuracy**
IV	572	0.84±6.29%
OOV	159	0.71±2.94%
Total	731	0.82±4.29%

Table 4: Test set performance on the metaphor detection task.

alization potential to unseen words. Unlike prior experiments, we detect a high variability in the results between repetitions; one model instance has a moderate performance that does not differ between the two subsets of the test set, whereas another achieves a near-perfect score on the IV subset while being barely above the random baseline in the OOV subset.

4 Related Work

NLI is widely considered one of the core problems towards natural language understanding, with a plethora of evaluation suites (Bowman et al., 2015; Conneau et al., 2018; Wang et al., 2018, 2019; Nie et al., 2020) which continue to pose significant challenge for current state-of-the-art models (Glockner et al., 2018; Talman and Chatzikyriakidis, 2019; Belinkov et al., 2019; McCoy et al., 2019; Richardson et al., 2020, *inter alia*). Like GLUE and SuperGlue, our inference examples come packed with linguistic tags to facilitate diagnostic analysis. Unlike other datasets, our examples may specify more than one inference label, accounting for all possible sentence readings. At the time of writing, other than a fragment of XNLI (produced by automatic translation), the only NLI dataset for Greek we are aware of is by Amanaki et al. (2022) (which we adapt here to our format).

Sense repositories, i.e., mappings between words and sets of meanings are often framed as dictionary-like structures (Fellbaum, 1998; Navigli and Ponzetto, 2012). Our dataset stands out in providing both a definition and a collection of examples for each sense, allowing the incorporation of either or both into various possible tasks and model pipelines; we show three concrete examples of how this can be accomplished. The tasks obtained, namely words-in-context, sense selection and metaphor detection, are of prime importance for the experimental validation of the lexical semantic capacities of language processing systems (Ma et al., 2021; Zhang and Liu, 2023; Choi et al., 2021; Sengupta et al., 2022; Luo et al., 2023). To the best of our knowledge, this is the first dataset of its kind, and among the first lexical resources for Greek in general.

5 Conclusions and Future Work

Our vision is that of an open-source, community-owned, dynamically adapted, gold-standard suite that enables the linguistically conscious evaluation of the capacities of Greek language models. We have presented four novel tasks and corresponding baselines towards that goal. While our results aren't directly comparable to existing benchmarks, they do highlight the significant challenge our tasks present. This underscores the urgency for accelerated progress within the Greek NLP ecosystem to stay aligned with contemporary mainstream research.

Pending community feedback, we hope to enrich the existing datasets by scaling them up, correcting possible artifacts and extending the language domain with regional and dialectal variations. Possible tasks that we would like the project to eventually incorporate include gender bias detection, paraphrase identification, and natural language inference with explanations, among others. We are curious to continue experimenting with ways to utilize traditional resources, and exploring their potential as dataset generators for under-resourced languages in conjunction with large language models.

Limitations

The NLI dataset's limited size renders it inadequate as a comprehensive resource for training and evaluating NLI systems from scratch. Furthermore, the examples were crafted by the authors of this paper, who belong to a distinct demographic, unavoidably introducing our own cultural, sociopolitical, and linguistic biases. The focus is exclusively on standard modern Greek, omitting examples of regional or dialectal language use. Finally, while the tag set employed may provide valuable information, it offers only a coarse and incomplete summary of the full range of linguistic phenomena observed in the wild.

The lexical dataset, conversely, is not indicative of our opinions as authors; the source dictionary may contain language use that is outmoded or socially exclusive. The dataset structure is sufficient for us to extract the three tasks we have presented, but might prove lacking for more complex tasks (like tasks requiring hierarchical or clustered sense arrangements, for instance). Despite efforts to ensure semantic accuracy in every entry, sense, and example, occasional mistakes may have gone unnoticed. Users should approach the resource critically, keeping this in mind.

Regarding our baselines, we have experimented with only a single model. While we acknowledge this might entangle the effects of dataset difficulty and model robustness, we justify ourselves in refraining from experimenting with more models, since this is neither the prime concern of this paper, nor a practice that we necessarily agree with.

References

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, Charikleia Soupiona, Effrosyni Skoulataki, and Christina Stefanidou. 2022. Fine-grained entailment: Resources for Greek NLI and precise entailment. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 44–52, Marseille, France. European Language Resources Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iakovos Evdaimon, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou. 2023. GreekBART: The first pretrained Greek sequence-to-sequence model. *arXiv preprint arXiv:2304.00869*.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2022. WiC = TSV = WSD: On the equivalence of three semantic tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2478–2486, Seattle, United States. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-BERT: The Greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Haochen Luo, Yi Zhou, and Danushka Bollegala. 2023. Together we make sense–learning meta-sense embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2638–2651, Toronto, Canada. Association for Computational Linguistics.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. Improvements and extensions on metaphor detection. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 33–42.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 143–146, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 34, pages 8713–8721.

Meghdut Sengupta, Milad Alshomary, and Henning Wachsmuth. 2022. Back to the roots: Predicting the source domain of metaphors using contrastive learning. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 137–142.

Aarne Talman and Stergios Chatzikyriakidis. 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. Association for Computational Linguistics.

G Triantafyllides. 1998. Dictionary of standard modern Greek. *Institute for Modern Greek Studies of the Aristotle University of Thessaloniki*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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. *arXiv preprint arXiv:1804.07461*.

Zi Wang and Daniel Hershcovich. 2023. On evaluating multilingual compositional generalization with translated datasets. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1687, Toronto, Canada. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.