Neurosymbolic AI for Natural Language Inference in French : combining LLMs and theorem provers for semantic parsing and natural language reasoning

Maximos Skandalis
LIRMM
CNRS & University of Montpellier
Montpellier, France
maximos.skandalis@lirmm.fr

Lasha Abzianidze
Institute for Language Sciences
Utrecht University
Utrecht, the Netherlands
l.abzianidze@uu.nl

Richard Moot
LIRMM
CNRS & University of Montpellier
Montpellier, France
richard.moot@lirmm.fr

Christian Retoré
LIRMM
CNRS & University of Montpellier
Montpellier, France
christian.retore@lirmm.fr

Simon Robillard
LIRMM
CNRS & University of Montpellier
Montpellier, France
simon.robillard@lirmm.fr

Abstract

In this article, we describe the first comprehensive neurosymbolic pipeline for the task of Natural Language Inference (NLI) for French, with the synergy of Large Language Models (CamemBERT) and automated theorem provers (GrailLight, LangPro). LLMs prepare the input for GrailLight by tagging each token with Part-of-Speech and grammatical information based on the Type-Logical Grammar formalism. GrailLight then produces the lambda-terms given as input to the LangPro theorem prover, a tableau-based theorem prover for natural logic originally developed for English. Currently, the proposed system works on the French version of SICK dataset. The results obtained are comparable to the ones on the English and Dutch versions of SICK with the same LangPro theorem prover, and are better than the results of recent transformers on this specific dataset. Finally, we have identified ways to further improve the results obtained, such as giving access to the theorem prover to lexical knowledge via a knowledge base for French.

1 Introduction

In Natural Language Processing (NLP), the classification task of predicting, for a given pair of sentences, the correct label between two (entailment, not entailment) or, better, three (entailment, neutral, contradiction) given ones is conventionally called Natural Language Inference (NLI) or Recognising Textual Entailment (RTE).

Deep learning methods have proven effective for the task, with quickly improving performance over the last years. However, they lack explainability, and they might predict a correct inference label based on heuristics that has little to do with reasoning but heavily relying on the nature of the training datasets (McCoy et al., 2019; Gururangan et al., 2018; Poliak et al., 2018). On the other hand, symbolic methods include using theorem provers for rule-based reasoning between the two sentences provided. In this case, the input has to be clearly structured. To get the best of both worlds, neurosymbolic AI methods can be used, where deep learning methods can be leveraged to prepare the input by converting the sentences to their logical form for the theorem prover, which is then used for reasoning on the sentences and outputs its label prediction as well as the proof with the rules it applied to reach this prediction.

After having introduced the context of the task and of the methods adopted, the article follows the structure below:

- We present already conducted research, first for English (Section 2.1), then for French (Section 2.2), both on the NLI datasets and on the neurosymbolic methods for NLI (Section 2.3 for preparing the input, and 2.4 for the logical methods for NLI).
- Section 3 lists and describes the steps for using neurosymbolic methods for NLI in French, providing the first pipeline for such use for

The code for the paper's pipeline is available on github. The datasets are all available on github and on huggingface.

Proceedings of the 16th International Conference on Computational Semantics, pages 252–263
September 22-23, 2025,

French.

- In Section 4.2, we analyse the work of adapting the tools for the case of French, due to the interlinguistic syntactic differences between the source language of the NLI theorem prover (English) and the target language (French).

- Some next steps for further improvement are outlined in Section 5.

2 Related work

2.1 Datasets in English

Numerous datasets exist in English for the task of NLI, namely FraCaS (Cooper et al., 1996), RTE1-8 (Dagan et al., 2006) (Dzikovska et al., 2013), SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), XNLI (Conneau et al., 2018), BreakingNLI (Glockner et al., 2018), ANLI and NLI-style FEVER (Nie et al., 2020), LingNLI (Parrish et al., 2021), GQNLI (Cui et al., 2022), WANLI (Liu et al., 2022), SpaceNLI (Abzianidze et al., 2023), the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. HANS (McCoy et al., 2019) and MED (Yanaka et al., 2019a) have only two labels, entailment and non-entailment.

In particular for logical reasoning with the natural language, eSNLI (Camburu et al., 2018) also contains natural language explanations for every label attributed. Finally, HELP (Yanaka et al., 2019b), ProofWriter (Tafjord et al., 2021), and FOLIO (Han et al., 2024) include First-Order Logical formulas for the sentences provided.

2.2 Datasets for French

For the task of NLI in French, significantly less datasets are available, despite some recent releases.

Table 1 gives the number of sentence pairs per class, for all the NLI datasets available in French, the first one, in order of release time, being XNLI (Conneau et al., 2018), FraCaS-FR(Amblard et al., 2020), then DACCORD (Skandalis et al., 2023), RTE3-FR, GQNLI-FR, and SICK-FR (Skandalis et al., 2024).

Because of the underrepresentation of contradictions in the widely used NLI datasets, it was recently proposed by Skandalis et al. (2023, 2024) to also work specifically on the labels contradiction/non-contradiction, with a new dedicated 2-class dataset for French, called DACCORD.

Dataset		Entailment	Neutral	Contradiction
SICK-FR	train	1274	2524	641
	dev	143	281	71
	test	1404	2790	712
FraCaS-FR	test	204	98	33
RTE3-FR	dev	412	299	89
	test	410	318	72
GQNLI-FR	test	97	100	103
XNLI-FR	dev	830	830	830
	test	1670	1670	1670
DACCORD	Rus-Ukr war		215	257
	Covid-19		251	199
	Climate change		49	63

Table 1: Breakdown by label for NLI datasets for French

2.3 Lambda-term or FOL formula extraction

In order to obtain the lambda-terms corresponding to a natural language sentence, one needs to first tag the tokens of the sentence with grammatical information. Categorial grammars are suited by design to producing lambda terms. While Combinatory Categorial Grammars (Steedman, 2000) have often been used in this context — for English notably the C&C (Clark & Curran) Parser (Clark and Curran, 2007) and EasyCCG (Lewis and Steedman, 2014) — we choose to use Type-Logical Grammars (TLG) instead. Type-Logical Grammars have the advantage of being purely logical formalisms, where lambda-terms are obtained by the Curry-Howard isomorphism. More pragmatically, our supertag models have been trained on the TLGbank for French, which uses Type-Logical Grammars as well. After the supertagger assigns formulas to each word, a parser is used to find the most likely parse for the given supertags.

These parses are then converted either to Lambda Logical Forms (LLFs), via components such as LLFgen (Abzianidze, 2017) or ccg2lambda (Martínez-Gómez et al., 2016), or to FOL formulas, usually with the intermediate step of the DRS (Discourse Representation Structure) formalism (Bos, 2008; Le, 2020). Lambda Logical Forms are simply typed λ-terms built up from variables and constant lexical terms with the help of two operations, function application and λ-abstraction.

More recently, Olausson et al. (2023) used Starcoder+ (Li et al., 2023) directly for FOL formula generation. The problem with this solution is that, unlike English, there were no datasets with sentences and their corresponding FOL representation for French, thus LLMs have not been previously exposed to such a task for French, in order to be able to handle it in some way.

For French, there are two main models for lambda-term extraction: DeepGrail and GrailLight (Moot, 2017). DeepGrail consists of both a supertagger and a parser, and the DeepGrail supertagger has been designed to integrate seamlessly with GrailLight. We have chosen to combine the DeepGrail supertagger with the GrailLight parser because this combination is the easiest to extend to a multi-tagger, as we will show in Section 3.1.5.

2.4 Theorem provers for natural language

Different theorem provers have been used for reasoning on natural language, specifically English:

- Coq (Chatzikyriakidis, 2015; Chatzikyriakidis and Bernardy, 2019; Bernardy and Chatzikyriakidis, 2021; Mineshima et al., 2015; Martínez-Gómez et al., 2017);
- LangPro (Abzianidze, 2015, 2017);
- Vampire (Bos, 2009; Bjerva et al., 2014; Haruta et al., 2022);
- Agda (Bekki and Satoh, 2015; Zwanziger, 2019);
- Prover9 (Olausson et al., 2023): Prover9 (McCune, 2005) is a theorem prover that attempts to solve theorems by contradiction and Mace4 attempts to find a counter-example to theorems.

A summary of their use on NLI can be found in Table 2.

2.4.1 LangPro theorem prover

LangPro (Abzianidze, 2017) is an automated theorem prover for natural logic (Muskens, 2010). It is written in Prolog, and makes use of the analytic tableau proof method. LangPro needs CCG (Combinatory Categorial Grammar) derivations of the linguistic expressions in order to obtain Lambda Logical Forms (LLFs) from them via the LLFgen (LLF generator) component. Otherwise, lambda terms that follow the following BNF syntax are the native format for the LangPro theorem prover itself:

```
TERM = (tlp(pl_atom_for_token,
    pl_atom_for_lemma,
    pl_atom_for_POS_tag,
    pl_atom_for_chunking_tag,
    pl_atom_for_named_entity_tag), TYPE
     | ( TERM @ TERM , TYPE ) | (abst(
    VAR, TERM ), TYPE )
VAR = (pl_var, TYPE)
TYPE = TYPE → TYPE | primitive_TYPE |
    featured_TYPE
primitive_TYPE = pr | pp
featured_TYPE = n:FEAT | s:FEAT | np:
    FEAT
FEAT = pl_var | dcl | ng | nb | pss |
    thr | adj | b | to | pt | rm | num
    | expl
```

`n` is the featured type assigned to nouns, `np` the type assigned to noun phrases, and `s` the type assigned to sentences.

In order to establish a certain logical relation between one or more premises and a hypothesis, the natural tableau method systematically searches for a counterexample that would invalidate the relation. The relation is considered proven if no such counterexample can be constructed; otherwise, the relation is refuted.

3 Pipeline Setup

3.1 Obtaining the input for the NLI theorem prover

3.1.1 POS-tagging

GrailLight theorem prover, which is used for the proof generation step, accepts Part-of-Speech tags from the TreeTagger tagset[1]. These POS-tags are also used for the semantics inferences by LangPro.

For TreeTagger POS-tags, three tools have been identified, either the original TreeTagger (Schmid, 2013) (which is now outdated) with a Python wrapper[2] for convenience, RNNTagger (Schmid, 2019), or the POS-tagger of the ELMo/bi-LSTM version of DeepGrail (Moot, 2021), which uses the model from Che et al. (2018). The latter one proved to be the best performing for this task.

Table 3 provides details on the number of occurrences of each POS-tag at the token level for French SICK dataset, as well as their partial correspondence with the tags in MELt tagset.

3.1.2 CG-supertagging with DeepGrail

The more recent Transformer version of the DeepGrail supertagger[3] uses CamemBERT (Martin et al., 2020), itself a French version of RoBERTa (Liu et al., 2019), for token embeddings. It is trained on the French Type-Logical Treebank

[1] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html.
[2] https://treetaggerwrapper.readthedocs.io/en/latest.
[3] https://gitlab.irit.fr/pnria/global-helper/deepgrail_tagger.

System	Proof strategy	Logic	Prover	Semantic parser	Abduction	Arithmetic	Datasets covered
Mineshima et al. (2015)	Ad hoc tactics	HOL	Coq	CCG Parser (C&C)			FraCaS
Abzianidze (2015, 2017)	Tableau	Natural logic / HOL	LangPro	C&C, and EasyCCG, then LLFgen	✓		FraCaS, SICK
Martínez-Gómez et al. (2017)	Ad hoc tactics	FOL	Coq	C&C, and EasyCCG	✓		SICK
Chatzikyriakidis and Bernardy (2019), Bernardy and Chatzikyriakidis (2021)	Ad hoc tactics	HOL	Coq	Grammatical Framework		✓	FraCaS
Haruta et al. (2022)	Resolution	Typed FOL	Vampire	C&C, EasyCCG, and depccg	✓ (WordNet and VerbOcean)	✓	FraCaS, MED, SICK, HANS & CAD
Olausson et al. (2023)	Resolution/model building	FOL	Prover9/Mace4	LLM (StarCoder+, GPT 3.5, GPT 4)			FOLIO & ProofWriter

Table 2: Existing methods based on theorem provers for NLI on English datasets

Number of occurences	TreeTagger tags	MELt tags
50725	NOM	NN (NNS?)
35984	DET:ART	DT
24269	PRP	IN
20471	VER:pres	VB
9416	ADJ	JJ
5447	ADV	RB
3886	KON	CC
3394	PRP:det	
3388	PRO:PER	PRP
3201	VER:pper	VBN
1876	NUM	CD
1461	PRO:REL	WP
832	PRO:IND	
645	VER:infi	VB
636	VER:ppre	VB
581	DET:POS	PRP$
398	PUN	
139	NAM	NNP
29	ABR	
24	PRO	PRP
23	PRO:DEM	DT
21	VER:simp	VBD
18	VER:impf	
14	VER:futu	
2	VER:subp	
2	SYM	

Table 3: Occurrences of each POS-tag in French SICK dataset for TreeTagger POS-tags and MELt POS-tags.

(Moot, 2015) to produce supertags (type-logical formulas) for each word in a sentence. DeepGrail is a loose adaptation of the work of Kogkalidis et al. (2020) to French. The supertagger assigns the correct formula to a word 96,1% of the time.

3.1.3 Lemmatisation

There are three tool options for lemmatisation for French, namely spaCy (Honnibal et al., 2020), Stanza (Qi et al., 2020), or Lefff (Sagot, 2010). Lemmas do not have an impact on the lambda-term extraction step, but they do have on the reasoning step with LangPro at the end. After inspecting the lemmatisation output, we concluded that Stanza's lemmatiser is comparatively the best among the three. For example, both spaCy and Lefff mistakenly gave as lemma `luire` for the word `lui` in the phrase `derrière lui`. On the other hand, Stanza gives the disjunctive pronoun `lui` as lemma for the subject pronoun `il`, indicating maybe that it groups pronominal forms together.

3.1.4 Proof and lambda-term generation with GrailLight

GrailLight (Moot, 2017) is a supertag-factored chart parser for multimodal type-logical grammars. It outputs a natural deduction proof for the highest-probability sequence of formulas for which a proof exists. A lambda-term for this proof is obtained by the Curry-Howard isomorphism.

Finally, we convert GrailLight's output lambda-term to LangPro's native input format shown in Section 2.4.1.

3.1.5 Evaluating the pipeline and improving the coverage

Dataset		Total sentences	Number of sentences parsed	Percentage of the sentences parsed (%)	Number of sentences failed to be parsed
SICK-FR		19,680	18,294	**92,96**	1,386
FraCaS-FR	test	882	838	**95,01**	44
GQNLI-FR	test	703	667	**94,88**	36
RTE3-FR	test	1,828	1,496	81,84	332
	dev	1,959	1,593	81,32	366
XNLI-FR	test	10,409	8,128	78,09	2,281
	dev	5,151	3,956	76,8	1,195
DACCORD		2,341	1,773	75,74	568

Table 4: Parsing results per dataset with 1 formula per token

In order to improve coverage from GrailLight, we used the 2022 Transformer version of DeepGrail Supertagger as a base, adding the beta value assignment introduced by Clark and Curran (2004) and already included in the 2021 ELMo/bi-LSTM version of DeepGrail (Moot, 2021).

For $P(x_i)$: the probability of predicted formula x_i,

$x_{\text{best}} = \arg\max_x P(x)$: the formula with the highest predicted probability,

β: the beta value (a scalar between 0 and 1),

$T = \beta \cdot P(x_{\text{best}})$: the threshold probability.

DeepGrail includes in its output, for every token, all formulas x_i such that:

$$P(x_i) \geq \beta \cdot P(x_{\text{best}})$$

It is to be noted that the beta value is not important per se; what matters is the resulting average number of predicted formulas per token.

Thus, without changing the pipeline (ELMo/bi-LSTM DeepGrail for POS-tagging, Stanza for lemmatisation, CamemBERT DeepGrail for CG supertagging), but with the beta value set to 0.01 and 0.0001 now (instead of set to 1.0 as in Table 4, or before in Skandalis et al. (2025)), which gives exactly one prediction per token), the number and percentage of proofs generated by GrailLight (whether these proofs are correct or not) are improved (see Tables 5 and 6).

Dataset		Total sentences	Number of sentences parsed	Percentage of the sentences parsed (%)	Number of sentences failed to be parsed	Average number of formulas per token
SICK-FR		19,680	19,564	99,41	116	1,0618
FraCaS-FR	test	882	869	98,53	13	1,0819
GQNLI-FR	test	703	688	97,87	15	1,0562
RTE3-FR-FR	test	1,828	1,775	97,1	53	1,15
	dev	1,959	1,890	96,48	69	1,176
XNLI-FR	test	10,409	9,748	93,65	661	1,1807
	dev	5,151	4,824	93,65	327	1,1913
DACCORD		2,341	2,196	93,81	145	1,1978

Table 5: Parsing results and formula density per dataset for beta value set to 0,01

Dataset	Total sentences	Number of sentences parsed	Percentage of the sentences parsed (%)	Number of sentences failed to be parsed	Average number of formulas per token
SICK-FR	19,680	19,644	99,82	36	1,4157
FraCaS-FR	882	881	99,89	1	1,8624
GQNLI-FR	703	698	99,29	5	1,2444

Table 6: Parsing results and formula density per dataset for beta value set to 0,0001

For comparison, Abzianidze and Kogkalidis (2021) report 95,9% of the sentences parsed for the dutch version of SICK with the Neural proof nets model from Kogkalidis et al. (2020), and 98,1% with the dutch Alpino parser (van Noord and Malouf, 2001).

3.2 Using LangPro for NLI for French

The LangPro has been initially developed for English but later adapted to Dutch (Abzianidze and Kogkalidis, 2021). We follow the previous work and in a similar style adapt the theorem prover to French. The main idea of the adaptation is to make the French terms somewhat similar to English terms as LangPro already has inference rules specialized for the latter ones. Such approach prevents us from making inference rules that specialize for French function words such as determiners and connectives. A brief illustration of transforming French terms into English-like terms is given below for the SICK NLI problem 3514, where the terms use lemmas of the corresponding words and non-French function words are highlighted in red.

(3514) P-FR: Une femme danse
`a femme danser`
H-FR: Il n'y a pas de femme qui danse
`ne`$^{\text{NIL}}$ $\big(\lambda y.$ `no (who danser femme)` $(\lambda x.$ `be` x $y)\big)$ `there`
P-EN: A woman is dancing
`a woman (be dance)`
H-EN: There is no woman dancing
`no (who dance woman)` $(\lambda x.$ `be x there)`
Label: Contradiction

More details on the adaptation is provided in Section 4.2. The entire pipeline of the French neurosymbolic NLI is concisely visualised in Figure 1.

4 Score and discussion

4.1 Score

We first evaluated some recent Transformer models on the French and English versions of SICK dataset. The results can be seen in Table 7. All NLI Transformer models for French are, in general, trained on the machine-translated from English to French train subset of XNLI. Thus, the evaluation of the LLMs is done here in cross-domain settings.

Model	SICK-EN		SICK-FR	
	Accuracy	Precision	*Accuracy*	Precision
DistilmBERT$_{\text{Base-cased}}$	52	61,25	48,43	54,01
XLM-R$_{\text{Base}}$	-	-	49,86	61,22
CamemBERT$_{\text{Base, 3-class}}$	-	-	52,89	63,63
mDeBERTa-v3$_{\text{Base, XNLI}}$	57,34	67,36	59,09	64,43
mDeBERTa-v3$_{\text{Base, NLI-2mil7}}$	**68,3**	**68,9**	**66,94**	**66,76**
XLM-R$_{\text{Large}}$	53,12	64,57	54,81	63,08
CamemBERT$_{\text{Large, 3-class}}$	-	-	58,3	64,83

Table 7: Results of label prediction by Transformers on SICK-EN and SICK-FR

Table 8 reports the results currently obtained on SICK-FR with LangPro theorem prover, with abduction and without the use of a dedicated French Knowledge base. It also gives for comparison the final results on SICK-EN and SICK-NL as reported by Abzianidze and Kogkalidis (2021), with the same theorem prover.

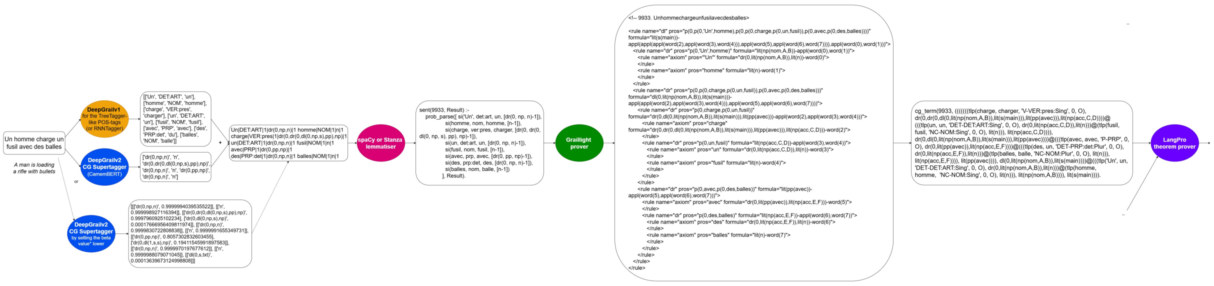

Figure 1: The pipeline for neurosymbolic NLI in French, with an example of conversion, which consists of the following steps: 1) POS-tagging and CG supertagging, 2) lemmatisation, 3) proof generation and lambda-term extraction, 4) theorem prover input.

Dataset		Accuracy	Precision
SICK-EN		84,4	94,3
SICK-NL (Abzianidze and Kogkalidis, 2021)		78,8	84,2
SICK-FR (present article)	test	71,1	96,8
	train-trial	76,9	98,6

Table 8: Precision and accuracy of LangPro for different languages

4.2 Handling inter-linguistic differences

Existential sentences with negation Historically in French, the word `ne` was the bearer of the sense of negation, and was followed by the word `point`, for emphasis. But nowadays, the negation is borne by the word `pas`, evolution of the word `point`. There are some occurrences where the word `ne` can appear without the `pas` to express the negation, but this is not with existential sentences. So for existential sentences, in order to align more easily the tree structures between `there exists/is no` and `il n'y a pas de`, we put together `pas de` as a quantifier, and correspond it to `no` as illustrated in 3514. While `ne` is still present in the corresponding term, it is marked with a specific NIL tag, indicating the semantic vacuousness for theorem proving.

Insert a WH-pronoun for VPs of type np→n→n To prove the contradiction such as the one in 3720, one needs to relate `épluche:np→np→s` to `épluchant:np→n→n` but it is difficult because of their different types. We convert `personne épluchant:np→n→n un oignon` into `personne WHICH:np→s→(n→n) épluchant:np→np→s un oignon`, which makes the connection between the verbs more transparent.

(3720) P-FR: Une personne épluche:np→np→s un oignon
H-FR: Il n'y a pas de personne épluchant:np→n→n un oignon
P-EN: A person is peeling an onion
H-EN: There is no person peeling an onion
Label: Contradiction

Attach remote "ne" to "personne" In sentences such as the premise in the example 4816, `ne` is renamed to `no` and attached to `personne`, so that the underlying logical form is `be (no (who ...) personne)` there, where closed-class words are replaced with English. With this, it is possible to prove the contradiction below.

(4816) P-FR: Il n'y a personne qui coupe un peu de gingembre
H-FR: Une personne coupe un peu de gingembre
P-EN: There is no person cutting some ginger
H-EN: A person is cutting some ginger
Label: Contradiction

Predicative adjectives In the English CCG, `be green` is analysed as `be:(np→s:adj)→np→s:dcl green:np→s:adj`, while in French TLG `be:(n→n)→np→s:dcl green:n→n` seems to be a preferred analysis. To accommodate the latter, the initial LangPro tableau rule empty_mod is extended, which discards `be:(n→n)→np→s:dcl`, and changes the type of `green` to `np→s:adj`. The analysis is intuitive, that's why it was accommodated in the inference rules rather than rewriting the French terms in the English style. This addition solves problems such as 3812 below:

(3812) P-FR: Une femme tranche un poivron qui est vert
H-FR: Une femme tranche un poivron vert
P-EN: A woman is slicing a pepper which is green
H-EN: A woman is slicing a green pepper
Label: Entailment

Normalise French terms Because of particularities of the chart rules, the French terms generated by GrailLight need not be in beta normal form.

(819) P-FR: Une personne en équipement de vélo est debout régulièrement en face de certaines montagnes
P-EN: A person in biking gear is standing steadily in front of some mountains
Label: Contradiction

The lambda-term for the example 819 above includes the subterm `(λ x. régulièrement(est debout x)) Une_personne_en_équipement_de_vélo`.

Before fixing any issues in the terms, first they are normalized.

Running abduction Abductive learning was introduced in LangPro by Abzianidze (2020). Abductive learning is run on the train and trial subparts of SICK, where LangPro has access to the gold inference labels and exploits them to learn useful lexical knowledge, i.e., relations over lexical items. In particular, LangPro induces the lexical knowledge that contributes to the proofs for entailment and contradiction problems. The learned lexical knowledge is later use to prove problems from the SICK-test subset.

Adding a knowledge base for access to lexical knowledge Results can be improved if we give access to the theorem prover to lexical relationships, such as hypernyms, synonyms, antonyms, geographical relations. For English, LangPro uses relations taken from WordNet 3.0 (Abzianidze, 2017). Knowledge bases, which could be used for this purpose for French, include the multilingual Babelnet (Navigli and Ponzetto, 2012), the monolingual French version of Wordnet WOLF (Sagot and Fišer, 2008), or JeuxDeMots (Lafourcade, 2007). Additional common sense knowledge, whose inclusion could be useful to test next, are listed in LoBue and Yates (2011).

As a first step here, we extracted the hypernyms (`isa`) and the antonyms from a 2013 version of JeuxDeMots, and converted them into Prolog format. This version contains 49.812 hypernyms, and 12.802 antonyms. Without further manipulation on the system, LangPro was able to prove some 52 additional problems from the train subset of SICK-FR with these relations. The example 5752 is one of these 52 cases, mentioning in `sys1` the prediction without access to the knowledge base, and in `sys2` the prediction that employs relations from JeuxDeMots.

(5752) P-FR: Le rhinocéros broute sur l'herbe
H-FR: L'animal broute sur l'herbe
P-EN: The rhino is grazing on the grass
H-EN: The animal is grazing on the grass
Label: Entailment
sys1: neutral
sys2: entailment, using isa(rhinocéros,animal)

We also extracted the same relations from a more recent version of JeuxDeMots (2024), amounting to 28.760.688 hypernyms and 131.813 antonyms, and plan on conducting tests with these versions, too.

Labels affected by translation Since this first version of SICK for French is machine-translated from English, some examples might need corrections in their translation after inspection, so that the initial label remains true.

(3181) P-FR: Un homme marche dans les bois
H-FR: L'homme ne marche pas dans les bois
P-EN: A man is trekking in the woods
H-EN: The man is not hiking in the woods
Label: Neutral

The example 3181 could be better translated, with an anglicism, as:

(3181) P-FR: Un homme fait un trek dans les bois
H-FR: L'homme ne fait pas de randonnée dans les bois
Label: Neutral

Finally, we applied manual corrections to the translations of certain sentences, the mistranslation of which may not impact the truth value of the label, or for which access to a knowledge base would now be needed in order for the label to remain truthful (e.g. *poivron vert* for green pepper, instead of *poivre vert* in the machine translation). These corrections are incorporated into the version of SICK-FR available on `github` and on `huggingface`.

5 Conclusion and perspectives

In this paper, we have presented the first combination of Transformers with automated theorem provers applied to the task of Natural Language Inference for French. The task of NLI with neurosymbolic methods can be split into two subparts: semantic parsing and natural language reasoning. The first one is necessary in order to convert the sentences to a form that can be processed by the theorem prover, that is, in the form of lambda terms or first-order logical formulae. In the case of French, to achieve this, one first needs to add Part-of-Speech and Type-Logical Grammar tags to the tokens of the sentences with the help of DeepGrail, then feed this to the Graillight logical parser. The LangPro theorem prover, that we chose here to use for the natural language reasoning, accepts lambda-terms as an input. We adapted it from English to French, mainly by aligning French linguistic structures to their equivalents in English, and by mapping words that can modify meaning to their English translations. The current performance of the model is promising, surpassing the performance of recent Transformer encoder models evaluated on the French SICK dataset. It is on par with the results obtained by LangPro on the English and Dutch versions of SICK, as long as more lexical knowledge is added for French as well. Finally, the present work also resulted in the first (NLI) datasets with sentences and their lambda-term representations available for French.

For the future, we plan to adapt and evaluate alternative semantic parsers, notably by using the DeepGrail parsers and by adapting Spindle (Kogkalidis et al., 2023) to generate lambda-terms for our French datasets. We also plan to extend the coverage of LangPro for French, so that it can handle FraCaS and GQNLI, as well. Finally, we aim at establishing another method based on a second theorem prover, for comparison reasons.

Acknowledgments

The research hereby presented was carried out with the financial support and approval of the French Ministry of Defence - Defence Innovation Agency (AID - DGA), to which we express our gratitude. This work was likewise supported by ICO, *Institut Cybersécurité d'Occitanie*, funded by *Région Occitanie*, France, which we would also like to thank. Finally, the first author was also funded by the Erasmus+ programme for a research stay at Utrecht University.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.

Lasha Abzianidze. 2017. LangPro: Natural language theorem prover. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 115–120, Copenhagen, Denmark. Association for Computational Linguistics.

Lasha Abzianidze. 2020. Learning as abduction: Trainable natural logic theorem prover for natural language inference. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 20–31, Barcelona, Spain (Online). Association for Computational Linguistics.

Lasha Abzianidze and Konstantinos Kogkalidis. 2021. A logic-based framework for natural language inference in dutch. *Computational Linguistics in the Netherlands Journal*, 11:35–58.

Lasha Abzianidze, Joost Zwarts, and Yoad Winter. 2023. SpaceNLI: Evaluating the consistency of predicting inferences in space. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 12–24, Nancy, France. Association for Computational Linguistics.

Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, and Sylvain Pogodalla. 2020. A French version of the FraCaS test suite. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5887–5895, Marseille, France. European Language Resources Association.

Daisuke Bekki and Miho Satoh. 2015. Calculating projections via type checking. In *ESSLLI proceedings of the TYTLES workshop on Type Theory and Lexical Semantics ESSLLI2015, Barcelona*.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2021. Applied temporal analysis: A complete run of the FraCaS test suite. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 11–20, Groningen, The Netherlands (online). Association for Computational Linguistics.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland. Association for Computational Linguistics.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286. College Publications.

Johan Bos. 2009. Applying automated deduction to natural language understanding. *Journal of Applied Logic*, 7(1):100–112. Special Issue: Empirically Successful Computerized Reasoning.

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

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Stergios Chatzikyriakidis. 2015. Natural language reasoning using coq: Interaction and automation. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 7–13, Caen, France. ATALA.

Stergios Chatzikyriakidis and Jean-Philippe Bernardy. 2019. A wide-coverage symbolic natural language inference system. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 298–303, Turku, Finland. Linköping University Electronic Press.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 103–110, Barcelona, Spain.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

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.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, Manfred Pinkal, David Milward, Massimo Poesio, Stephen Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Technical report, FraCaS: A Framework for Computational Semantics. FraCaS deliverable D16, 136 pages, also available by anonymous ftp from `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2022. Implementing natural language inference for comparatives. *Journal of Language Modelling*, 10(1):139–191.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303.

Konstantinos Kogkalidis, Michael Moortgat, and Richard Moot. 2020. Neural proof nets. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 26–40, Online. Association for Computational Linguistics.

Konstantinos Kogkalidis, Michael Moortgat, and Richard Moot. 2023. Spindle: Spinning raw text into lambda terms with graph attention. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 128–135.

Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand.

Ngoc Luyen Le. 2020. *French language DRS parsing*. Theses, Ecole nationale supérieure Mines-Télécom Atlantique.

Mike Lewis and Mark Steedman. 2014. A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar. Association for Computational Linguistics.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you!

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90, Berlin, Germany. Association for Computational Linguistics.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 710–720, Valencia, Spain. Association for Computational Linguistics.

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.

W. McCune. 2005. Prover9 and mace4. http://www.cs.unm.edu/~mccune/prover9/.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Richard Moot. 2015. A type-logical treebank for french. *Journal of Language Modelling*, 3(1):229–264.

Richard Moot. 2017. *The Grail Theorem Prover: Type Theory for Syntax and Semantics*, pages 247–277. Springer International Publishing, Cham.

Richard Moot. 2021. *Type-logical investigations: proof-theoretic, computational and linguistic aspects of modern type-logical grammars*. Accreditation to supervise research, Université Montpellier.

Reinhard Muskens. 2010. An analytic tableau system for natural logic. In *Logic, Language and Meaning*, pages 104–113, Berlin, Heidelberg. Springer Berlin Heidelberg.

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, Online. Association for Computational Linguistics.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco.

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, pages 154–164. Routledge.

Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 133–137, New York, NY, USA. Association for Computing Machinery.

Maximos Skandalis, Lasha Abzianidze, Richard Moot, and Simon Robillard. 2025. Hybrid AI with LLMs and Theorem Provers for Semantic Parsing and Natural Language Inference for French. FoMo 2025 - ELLIS Winter School on Foundation Models. Poster.

Maximos Skandalis, Richard Moot, Christian Retoré, and Simon Robillard. 2024. New datasets for automatic detection of textual entailment and of contradictions between sentences in French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12173–12186, Torino, Italia. ELRA and ICCL.

Maximos Skandalis, Richard Moot, and Simon Robillard. 2023. DACCORD : un jeu de données pour la détection automatique d'énonCés COntRaDictoires en français. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 285–297, Paris, France. ATALA.

Mark Steedman. 2000. *The syntactic process*. MIT Press, Cambridge, MA, USA.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

G.J.M. van Noord and R. Malouf. 2001. Alpino: Wide-coverage computational analysis of dutch. In *Computational Linguistics in the Netherlands 2000*, LANGUAGE AND COMPUTERS : STUDIES IN PRACTICAL LINGUISTICS, pages 45–59. Rodopi. Joke; 11th Conference on Computational Linguistics in the Netherlands ; Conference date: 03-11-2000.

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*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Colin Zwanziger. 2019. Dependently-typed Montague semantics in the proof assistant agda-flat. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 40–49, Toronto, Canada. Association for Computational Linguistics.