# Does Logic-based Reasoning Work for Dutch?

**Lasha Abzianidze** and **Konstantinos Kogkalidis**
Utrecht Institute of Linguistics OTS, Utrecht University
{l.abzianidze,k.kogkalidis}@uu.nl

## 1 Introduction

Natural Language Inference (NLI) is a hallmark task to evaluate an NLP system's capacity for natural language understanding. In recent years, the surge in the performance and accessibility of neural language models has led to a surplus of generic treatments to NLI, revolving around the "pre-train & fine-tune" paradigm. Despite their state-of-the-art results, such models suffer from a series of downsides, opaqueness and brittleness being the most notable. When presented with a query, a task-agnostic neural model will always produce an answer, but offer no transparency in its decision procedure. Furthermore, the performance of the aforementioned models often degrades on the out-of-the-distribution samples. Recent works found that the deep learning models high performance is partially due to exploiting annotation artifacts. In stark contrast to the above, logic-based approaches for NLI boast transparency, robustness and formal rigor, at the cost of an (often marginal) drop in performance. Logic-based systems offer not just a prediction, but the full explanation behind it, allowing a deeper inspection of their inner workings that goes beyond mere quantitative comparisons.

In this work, we utilize the recently released Dutch translation of the SICK dataset (Wijnholds and Moortgat, 2021) as an experimental test bed for the cross-lingual application of the Natural Tableau-based theorem prover LangPro (Abzianidze, 2017). We obtain syntactic analyses for the dataset in the form of Typelogical Grammar (TLG) derivations from two wide-coverage parsers: ALPINO (rule-based) and Neural Proof Nets (NPN, neurosymbolic). After converting them to logical forms and combining them with lexical relations extracted from the Open Dutch WordNet (ODWN, Postma et al., 2016), we use them to train LangPro in a parameter-free abductive setting.

Our contributions can be summarized as follows:

- The first logic-based NLI system for Dutch demonstrates promising results on SICK-NL;
- Abduction discovers knowledge that is complementary to ODWN;
- Our experiments serve as external evaluation and comparison between ALPINO and NPN;
- Semi-automatically translation of an NLI dataset might make it more challenging.

## 2 Methodology

We use the higher-order logic-based theorem prover LangPro to solve Dutch inference problems. LangPro was designed for English. It uses English parsing models of Combinatory Categorial Grammar (CCG, Steedman, 2000) and the Princeton Wordnet of English (Miller, 1995). Dutch localization of LangPro requires at least two language-specific key components: a robust syntactic parser with a transparent syntax-semantic interface, and a lexical knowledge database. We find ready solutions for both in existing research. For our abstract syntactic representations, we employ TLG derivations in the Intuitionistic Linear Logic flavor, which carry the benefit of being both highly refined and neutral with respect to the application domain, allowing a smooth transition to logical forms. We obtain analyses from two distinct sources: ALPINO (Bouma et al., 2001), a graph parser based on a maximum entropy disambiguation model and a hand-written rule system, and NPN (Kogkalidis et al., 2020a) a domain-specific neurosymbolic parser that directly translates natural language utterances to TLG proofs & terms. We homogenize the two parsers' outputs by transforming ALPINO graphs to TLG derivations using the conversion script of Kogkalidis et al. (2020b). For the lexical knowledge component, we employ ODWN, an open-access resource derived from the

$$\textbf{large}_{\mathrm{np,np}}\ \big(\textbf{brown}_{\mathrm{np,np}}\ (\textbf{a}_{\mathrm{n,np}}\ \textbf{dog}_{\mathrm{n}})\big) \rightsquigarrow \textbf{a}_{\mathrm{n,np}}\ \big(\textbf{large}_{\mathrm{n,n}}\ (\textbf{brown}_{\mathrm{n,n}}\ \textbf{dog}_{\mathrm{n}})\big) \quad (1)$$

$$\textbf{and}\ \big(\lambda x.\,\textbf{brown}(x\ \textbf{dog})\big)\big(\lambda y.\,\textbf{black}(y\ \textbf{dog})\big)\ \textbf{no} \rightsquigarrow \textbf{and}_{\mathrm{np,np,np}}\ \big(\textbf{no}\ (\textbf{brown}\ \textbf{dog})\big)\ \big(\textbf{no}\ (\textbf{black}\ \textbf{dog})\big) \quad (2)$$

$$\textbf{cut}_{\mathrm{pp,n,np,s}}\ (\textbf{in}_{\mathrm{n,pp}}\ \textbf{slice}_{\mathrm{n}})\ \textbf{meat}_{\mathrm{n}} \rightsquigarrow \textbf{cut}_{\mathrm{pp,np,np,s}}\ (\textbf{in}_{\mathrm{np,pp}}\ \{\textbf{slice}_{\mathrm{n}}\}_{\mathrm{np}})\ \{\textbf{meat}_{\mathrm{n}}\}_{\mathrm{np}} \quad (3)$$

proprietary Cornetto (Vossen et al., 2008), and standardize its entries to make compatible with the Princeton WordNet 3.0 standards.

We convert TLG terms into Lambda Logical Forms (LLFs). The first part of the conversion strips modality information from TLG types and terms and carries out possible $\beta\eta$-reductions. The obtained simply typed lambda terms are later modified with seven rewriting rules to make the terms semantically adequate as expected by the prover. The examples of the fixed structures (presented as the English literal translations) include fixing the order of nominal modifiers and a determiner (1), distributing determiners over conjuncts (2), and decreasing the number of lexical categories via type-changing rules (3).

Adaptation of LangPro to Dutch data was carried out on the development part (500 problems) of SICK-NL. In addition to designing the term rewriting rules, this phase includes mapping semantically-transparent Dutch function words (e.g., negation, determiners, connectives, and auxiliary verbs, but not prepositions(!)) to canonical constant terms, relevant for the tableau inference rules. We also map Dutch-specific POS tags of lexical terms to Penn Treebank-like tags.

## 3 Experiments & Results

We run experiments with LangPro using combinations of a parser, ALPINO or NPN, and a POS tagger, ALPINO or spaCy (Honnibal et al., 2020). The development and training parts (in total 5K problems) of SICK-NL were used for abductive learning (Abzianidze, 2020) of lexical knowledge. The results in Table 1 show comparable performances of the parsers and a significant impact of POS tagging on the results. As demonstrated in the previous works, we aggregate proofs for all combinations to achieve the best result (i.e., LangPro NL2×2). While, unsurprisingly, the transformer-based models outperform LangPro, the latter still proves problems that are misclassified by them, especially the entailment ones (see Table 2).

The adaptation phase and error analysis revealed that SICK-NL is more challenging than its original English version due to the semi-automatic translation of the dataset. For example, the problem ⟨*A man is trekking in the woods*, *The man is not hiking in the woods*⟩ with a gold but arguable NEUTRAL label is translated into a clear CONTRADICTION pair as both *trekking* and *hiking* are mapped to the same word *wandelen*. Several phrases shared by the English premise and hypothesis are translated into different phrases apparently due to the context sensitivity of the machine transition model. For instance, *drawing a picture* ↦ *een tekening maakt* | *tekent een foto*, *dirt bike race* ↦ *crossmotorwedstrijd* | *crossmotorrace*. Such misalignments of phrases make the problems more challenging.

An example of a useful relation learned by the abduction, and not found in ODWN, is the synonymy of the senses of *lopen* and *rennen*. The relation is often crucial for theorem proving because of the context-sensitive translation of *run*. Other found relations include **halter**/*dumbbell* ⊑ **gewicht**/*weight*, **pizza** ⊑ **voedsel**/*food*, and **leeg**/*empty* | **vol**/*full*.

| Parser / POS | $T\epsilon$ | $T\alpha$:$T\epsilon$ | $T\alpha$:$E\epsilon$ |
|---|---|---|---|
| Alpino / Alpino | 72.7 | 82.0 | 75.9 |
| Alpino / spaCy | 74.8 | 84.3 | 77.6 |
| NPN / Alpino | 72.0 | 80.6 | 74.3 |
| NPN / spaCy | 74.3 | 83.4 | 76.4 |
| **LangPro NL2×2** | 76.0 | 85.8 | 78.7 |
| LangPro EN3 | 83.2 | 91.1 | 84.4 |

Table 1: Accuracy scores (%) of the experiments. The abbreviations used: T (SICK-NL-train-dev, 5K problems), E (SICK-NL-evaluation, 4927 problems), $X\alpha$ (training on X with abduction), $X\epsilon$ (evaluating on X).

| Models | All ±Δ | Ent ±Δ | Cont ±Δ |
|---|---|---|---|
| LP NL2×2 | 78.7 | 50.6 | 66.3 |
| BERTje | 82.0 −0.3 | 86.2 +2.0 | 86.7 +0.8 |
| mBERT | 79.9 +0.7 | 79.0 +4.7 | 81.9 +3.1 |
| RobBert | 81.7 +0.9 | 76.9 +6.4 | 85.3 +1.1 |

Table 2: Comparison of transformer-based models and LangPro on SICK-NL-evaluation, based on all and label-specific problems. ±Δ shows the difference when the predictions of the transformer-based models are replaced only with LangPro's ENT./CONT. predictions.

## References


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.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of dutch. In *Computational linguistics in the Netherlands 2000*, pages 45–59. Brill Rodopi.

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

Konstantinos Kogkalidis, Michael Moortgat, and Richard Moot. 2020a. 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. 2020b. Æthel: Automatically extracted typelogical derivations for dutch. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5259–5268, Marseille, France. European Language Resources Association.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA, USA.

Piek Vossen, Isa Maks, Roxane Segers, and Hennie VanderVliet. 2008. Integrating lexical units, synsets and ontology in the cornetto database. In *LREC*.

Gijs Wijnholds and Michael Moortgat. 2021. Sicknl: A dataset for dutch natural language inference.