Linguistic Theories Coincide with Misformalization in Temporal Logic

Colin S. Gordon
Department of Computer Science, Drexel University
Philadelphia, PA, USA
csgordon@drexel.edu

***Abstract*—One of the key challenges in using formal methods is producing accurate formalizations of natural language requirements, as providing incorrect formalizations may miss bugs or even codify their existence. Yet despite this critical role, recent studies have revealed that even experienced experts make mistakes when formalizing relatively simple specifications in Linear Temporal Logic (LTL). We analyze the data from one recent study from the perspective of linguistic phenomena that enrich what is said with additional meaning. We find that misunderstanding whether and when to formalize these phenomena could impact nearly half of novice mistakes and most expert mistakes in the dataset. We conclude that further study of the relationship between natural language specifications and these specific phenomena has potential to reduce misformalizations.**

***Index Terms*—linguistics, pragmatics, temporal logic, formal specification**

I. Introduction

Formal methods have long been a promising approach to improving software correctness, and after years of deployment in safety-critical domains or high cost-to-repair domains (such as hardware), are seeing a resurgence of interest in parts of the software industry [1], [2]. However most formal methods require a formal specification of the properties to check. These specifications are typically obtained by manual translation of natural language specifications into a tool-specific formalism (LLM assistance is briefly discussed in Section IV). This translation is critical: formalizing a property incorrectly may codify that some misbehaviour *should* exist (instead of shouldn't) or may under-constrain the system's behaviour; either results in missing defects.

Moreover it is error-prone. Greenman et al. have recently conducted studies [3], [4] analyzing the sorts of mistakes made even with small, single-sentence specifications — by both students and experts — when translating into Linear Temporal Logic (LTL) [5]. Their detailed analysis breaks down the specific symptoms of mistakes which appear in misformalizations. However, they only characterize possible causes of mistakes based on reasonable interpretation as experienced formal methods researchers and teachers, leaving aside subtleties linguists have studied in how natural languages express time.

In linguistics, the expression of temporal ordering has emerged as one of the most subtle aspects of natural language meaning. Therefore, we hypothesize that many misformalizations of temporal properties are likely due to the subtleties of general natural language interpretation, rather than confusions about the target formalism. We previously posited [6] that some mistakes may be due to *pragmatic* factors — those related to the implicit intended meaning carried by natural language, which goes beyond what is explicitly stated — and that disagreements between how time naturally advances in natural language narratives and does not necessarily advance in temporal logics may explain other mistakes [7]. However those claims were not grounded in careful data analysis. This paper tests those theories by analyzing mistake data from Greenman et al.'s earlier English-to-LTL translation experiments, and comparing those factors to the incorrect answers provided by their study participants. We find that nearly half of student mistakes (12/25) and a solid majority of expert mistakes (12/15) are related to specific linguistic factors discussed posited in earlier work (explained below). This means that these well-studied aspects of natural language are likely significant factors in misformalization with LTL, and more user awareness of how these aspects of natural language function may lead to better formalization, instruction, and tool support.

II. Pragmatics, Narrative Time, and Habituals

Pragmatics is a subfield of linguistics addressing the gap between what is explicitly said and what meaning is intended by the speaker to be communicated. A classic example is

I got on my horse and I rode into the sunset. (1)

Here the word "and" strictly speaking denotes logical conjunction — which does not directly imply anything about the temporal ordering of conjuncts. But (1) is understood to mean that the speaker/writer *first* got onto the horse and only after that rode into the sunset, as opposed to first riding off in a go-kart and then getting on a horse at the destination. This is an example *conjunction buttressing* [8] or *asymmetric conjunction*, where the conjunction "and" picks up additional intended meanings (here "and then") that affect the two conjuncts differently. But this is just an example — any meaning that an author intends to communicate, but which is not explicitly stated, is considered an *implicature* [9]. For our purposes we need only scratch the surface of this rich topic [8]. But one key concept is the standard *test* for implicature: a part of an interpretation is a pragmatic implicature — an enrichment beyond what is explicitly stated — if it can be *cancelled* by additional context, as in (2):

I got on my horse, and earlier I rode into the sunset. (2)

2643-1572/25/$31.00 ©2025 IEEE
DOI 10.1109/ASE63991.2025.00329

"and" in (2) cannot inherently impose temporal ordering on its conjuncts in (1), or (2) would read as contradictory. We can (and do) use such tests to check for pragmatic phenomena.

Narrative time is a related pragmatic concept [10]. In (3) we understand events to have occurred in the exact order written.

I wrote the formal specification, ran the model checker, and fixed some bugs. (3)

This is also a cancellable implicature [11], [12]: if the sentence were extended by "...but not in that order" the ordering assumption is gone. This kind of temporal implicature is pervasive in every-day language usage, but temporal logics require any ordering to be directly expressed.

Habitual sentences are another clearly-related topic in linguistic theory. Habitual sentences characterize a recurring "habit" of a subject over time, encompassing many independent instances of some event of interest [13]. Many sentences have both habitual and non-habitual readings, and *in English* are systematically ambiguous in isolation as to which interpretation they receive. Declerck [14, p. 144] nicely explains this ambiguity by noting that "each of these instances [of the habit] can be described in terms of the same sentence as is used to refer to the habit as a whole." Consider (4):

The door will open when you approach (4)

This could be read as the door opening *every* time someone approaches the door, or as only applying to one specific (future) occasion of approaching the door.

Clearly these temporal aspects of linguistics are *potentially* relevant to formalizing temporal specifications. We can actively construct examples where formalizing an implicature or not, or a habitual or episodic interpretation, is the difference between correct and incorrect formalizations (observable in misuse or absence of certain temporal operators in LTL). Our analysis in Section III establishes that these specific phenomena are not merely plausible, but likely sources of mistakes observed in user studies.

III. (Re)Analysis of Greenman et al.'s 2022 Data

Greenman et al. [3] investigate the mistakes made by a number of participants when working with Linear Temporal Logic (LTL) [5]. Across 4 rounds of experiments, they investigate the understanding of the relationship between an LTL formula and its models, LTL-to-English translation, and English-to-LTL translation. The first two tasks are certainly relevant to English-to-LTL translation (e.g., a formalizer will likely make mistakes with an operators whose logical semantics are misunderstood). But we focus on the latter, English-to-LTL task which directly asks participants to formalize their understanding of the English.

Because Greenman et al. report changing wording after the earlier rounds to fix what the authors recognized as possibly-confusing wording, we consider only the 3rd and 4th rounds of experimental data, whose English can be considered polished, high quality natural language specifications: they were written by experts with clarity as an explicit goal, and refined for clarity based on previously-observed ambiguities. Our analysis shows that even then, the specific linguistic phenomena identified in Section II can explain many mistakes. Of note, round 3 participants were all active researchers with LTL experience, while round 4 participants (like rounds 1 and 2) were students who had been taught LTL in the context of a formal methods course.

Both groups were asked to translate five sentences to LTL:

1) Whenever the Red light is on, it is off in the next state and on again in the state after that
2) The Red light is on in exactly one state, but not necessarily the first state
3) The Red light cannot stay on for three states in a row
4) Whenever the red light is on, the Blue light will be on then or at some point in the future
5) The Red light is on for zero or more states, and then turns off and remains off in the future.

It is critical to note that our reanalysis is *not* simply ad hoc nit-picking of Greenman et al.'s evaluation instruments. Instead, we are evaluating the relationship between mistakes in formalizing these instruments and established linguistic theories that might explain *why* certain mistakes are made. Most real-world natural language descriptions of temporal software behaviour will not be subjected to the level of care that Greenman et al. have already exercised in refining these instruments. So if we can identify causes of systematic patterns in mistakes that exist in this data (which has been carefully curated by experts to avoid confusion and ambiguity) then the same causes are almost certain to play a role when formalizing less polished specifications in the real world, and work in language instruction and natural language processing on these specific phenomena are likely to reduce mistakes when incorporated into LTL instruction and tools.

A. Analysis Approach

We obtained Greenman et al.'s publicly shared (clearly-coded) data, which are available as a spreadsheet [15]. For each *incorrect* English-to-LTL translation in rounds 3 and 4, we (the author) examined the original sentence, correct translation, and incorrect translation together, and evaluated whether the incorrect translation could arise as a meaning if different pragmatic implicatures or habitual resolutions were made vs. the correct meaning. We began with no fixed set of possible causes, but open-minded inspection of the mistakes quickly converged on three phenomena. After coding each mistaken answer from rounds 3 and 4 for which phenomena appeared connected and also for the direction of misformalization, a second round of review was employed several weeks later to confirm, and as a result one category was split into two (and one habitual label was removed). For each category a participant might have formalized an implicature that was plausible but not intended, or a participant may have failed to formalize an implicature which was intended. Some responses were given multiple tags. Tags were only applied if we could affirmatively state a pragmatics/habitual explanation. Concepts that applied to only one response were discarded.

B. Analysis Results

Our findings are summarized in Table I. For brevity, rather than enumerate sentence by sentence results, we explain the resulting categories and discuss trends and interesting phenomena in the results. Our data is publicly available [16].

a) Habitual Misformalization: This category corresponds to mistakes involving confusion over whether some claim was meant to hold generally (i.e., a habitual interpretation true throughout time) or only at a single moment in time (i.e., an episodic / one-time interpretation). In practice these mistakes are almost always omitting a top-level "always," as in the participant writing $red \Rightarrow (\bigcirc \neg red \wedge \bigcirc \bigcirc red)$ instead of $\Box(red \Rightarrow (\bigcirc \neg red \wedge \bigcirc \bigcirc red))$ for sentence 1.

b) Conjunction Buttressing: This category contains mistakes in whether "and" should receive a temporal or atemporal interpretation. This usually appears as cases where adding an extra "next-state" ($\bigcirc$) operator around the right conjunct of a logical "and," where when the ordering of conjuncts was the same in English, would make the answer more correct. For example, one participant wrote $(\neg red) \mathsf{U} (red \wedge (\Box(\neg red)))$ for sentence 2 instead of $(\neg red) \mathsf{U} (red \wedge (\bigcirc \Box(\neg red)))$. That participant's mistake could be explained by assuming that right conjuncts in the logic are silently shifted later in time as in 1 (which is what the extra $\bigcirc$ in the solution accomplishes), as in classic conjunction buttressing implicatures in English. This was observed only in experts.

c) Implication Buttressing: This category contains mistakes in whether a conditional should receive a temporal interpretation or not. Similar to conjunction buttressing, this corresponds to a misformalization that would be correct if the conclusion of an implication were shifted later in time than the antecedent, as often occurs in English. For example, one participant wrote $((\neg red) \mathsf{U} red) \wedge \Box(red \Rightarrow \Box(\neg red))$ for sentence 2 instead of $((\neg red) \mathsf{U} red) \wedge \Box(red \Rightarrow \bigcirc \Box(\neg red))$ (the latter is correct). This mistake was observed only with student participants.

d) Narrative Time: This category includes mistakes involving incorrect assumptions about some notion of current time advancing throughout a specification (whether the advancement was "assumed" over English or LTL). This was observed twice, both with experts. One example for sentence 1 is:

$$(\diamond(red)) \Rightarrow (\bigcirc(\neg red \wedge (\bigcirc(red))))$$

Setting aside the missing "always" around the whole formula (this misformalization also gets the habitual semantics incorrect, see Section III-C) this would otherwise be correct if the right side of an implication was evaluated at the *first moment in time witnessing truth of the antecedent*, which is how the actual English is evaluated under theories of narrative time advancing in discourse [17]. That is, this mistake could be explained by the participant expecting additional implicatures to further enrich the meaning of the logical formula written. Of note, most participants (including the one making this mistake) responded using an "anglicized" LTL, where logical implication was literally presented by an infix "implies" keyword.

This category was split off from the previous during the review round. They were one category in the initial review because both categories' manifestations include incorrect temporal shifting related to conditional implication, but the linguistic phenomenon at play differs between the two categories.

C. Comparing Our Categorization to Greenman et al.'s

As with Greenman et al., some responses have multiple tags. Our hygienics label is closely related to Greenman et al.'s *Implicit Global* (IG) error category.[1] Every mistake Greenman et al. labeled IG we labeled hygienic. We labeled one additional mistake hygienic: the narrative time example above, which their public dataset labeled with a question mark. We believe that participant formalized what Declerck calls the "attributive definite" interpretation of the sentence [14], in which the "whenever" asserts the existence of a single unknown time,[2] rather than the intended habitual interpretation. Greenman et al.'s *Bad State Index* category is closely associated with our buttressing categories — technically the fixes change the index (time) where part of the formula is evaluated — but categorizes the symptom, not plausible causes. Greenman et al. also have a *Reasonable Variant* category, which is a catch-all for participants coming up with unintended interpretations and formalizing them correctly. They tagged only one example this way in the re-analyzed data: a translation of a habitual sentence as a one-time occurrence (which we labelled as habitual-related). This category is broader than and narrower than our categories in different ways that make comparison awkward: it is broader by allowing any alternative interpretation regardless of likelihood or cause (we restrict ourselves to several phenomena grounded in established linguistic theories); and it is narrower by only allowing *fully correct* formalizations of the alternatives (we applied labels if they explained even part of a mistranslation clearly).

D. Limitations

Our reanalysis is based on only the contents of Greenman et al.'s original paper [3], and their publicly-shared anonymous dataset. Thus while a substantial portion of participant mistakes *could* be explained by interactions between pragmatics and the formalization process, the data alone do not rule out the possibility that some of the mistakes we are associating with pragmatics were merely random slip-ups. However the fact that the mistakes associated with possible pragmatic factors are also some of the most prevalent — and co-occur in a number of responses — suggests that pragmatic phenomena do play some substantive role in LTL misformalization.

Another possible limitation relates to natural language acquisition. Pragmatic implicatures are some of the most challenging aspects of learning a new language [18]. If some of Greenman et al.'s participants were not native English speakers, it is possible that some pragmatics-derived mistakes are more strongly related to the subtlety of learning pragmatic implicatures in a 2nd (or 3rd, or later) language. They do not report on native speaker status of their participants, but given the demographic make-ups of both the institution of

[1] The "always" operator of LTL is sometimes called the "globally" operator.
[2] Consider "Whenever Tom commits that fix I must merge."

TABLE I
SUMMARY OF PRAGMATIC IMPLICATURE ANALYSIS OF MISTAKES. HABITUALITY, "CONJ.", "NARR.", AND "IMPL." ARE THE FOUR CATEGORIES; OMITTED COLUMNS WERE NOT OBSERVED. "UNDER" REFERS TO UNDER-FORMALIZING AN IMPLICATURE THAT WAS EXPECTED IN THE FORMAL SPECIFICATION. "OVER" REFERS TO FORMALIZING ADDITIONAL IMPLICATURES WHICH ARE PLAUSIBLE BUT WERE NOT INTENDED.

Sentence	**Expert Mistakes (Exp. 3)**						**Student Mistakes (Exp. 4)**			
	Habituality	***Conj.***	***Narr.***	***Under***	***Over***	***Total***	***Habituality***	***Impl.***	***Under***	***Total***
1	2	0	1	2	0	2	1	0	1	1
2	2	3	0	3	0	3	2	2	4	4
3	1	0	0	1	0	1	2	0	2	2
4	2	0	0	2	0	2	2	0	2	2
5	1	0	1	3	1	4	3	0	3	3

the undergraduate participants and the overall demographics of computer science academia (for experts), it is likely that some participants either learned a different world English (with sometimes-differing pragmatic implicatures [19]), or learned English later in life and therefore may not have (yet) internalized all of the many pragmatic implicatures expected in English. More research is necessary to tease apart these factors. Real-world settings also include non-native speakers of the working language, so there is value in further investigating the relationship between pragmatics and LTL formalization, even if the benefits might be unevenly distributed among formal methods practitioners.

Our analysis was focused on the widely-known, heavily-studied phenomena described in Section II. It is possible that other well-established linguistic phenomena are also correlated with these mistakes, or are correlated with the mistakes that we did not identify correlations with. Based on the correlations uncovered, consideration of other linguistic phenomena are also worthwhile future investigations.

IV. RELATED WORK

Research into easier and/or more reliable translation from natural language (always English in these studies) to logic is a classic idea. Gordon and Matskevich [20] survey linguistics-based approaches for non-temporal software specifications. Brunello et al. [21] survey automated translation for temporal specifications up through 2019. While some temporal specification work draws on the linguistics literature (notably Dzifcak et al. [22], and Nelken and Francez [23]), most of that work (and more modern work) does not attempt to investigate the specific aspects of the translation process that are challenging for humans (though all of it assumes the formalization process is difficult and error-prone). Greenman et al. [3], [4] are one of the only groups to perform user studies to characterize specific mistakes, but their analysis is not grounded in linguistic theories of how natural language encodes time information. We previously suggested [6], [7] that linguistic studies of pragmatics might offer insight into these mistakes based on Greenman et al.'s results and paper examples, but this paper is the first to systematically analyze LTL misformalization data with an eye towards how established linguistic theories may contribute to mistakes. Others have studied the human challenges in formalization systematically (notably Finney [24]). Vinter [25] hypothesized linguistic interference in specification formalization and carried out a large user study of drawing incorrect logical conclusions from Z specifications, based on 1970s psychology work predating the modern linguistic theories of pragmatics we consider. Other researchers have studied the linguistic features of how developers use language in other natural language software engineering artifacts [26], [27].

An obvious question today is how Large Language Models (LLMs) fare on this task. The general trends are similar to other programming-related tasks: they can simplify some parts of translation, but the unpredictable (and for temporal logic, sparse) nature of the underlying probability distribution makes manual audit essential [28]–[30]. The fact that even experts still make key mistakes in LTL formalization suggest that code review of LTL specifications is likely more error-prone than typical code review of LLM output. It seems reasonable to assume that difficulties in human review of automatically-generated formalizations will resemble difficulties in human translation, in which case further investigation of the pragmatic lens on LTL formalization can yield benefits there as well.

V. CONCLUSIONS, IMPLICATIONS, AND FUTURE WORK

We have demonstrated that in the data from Greenman et al.'s earlier experiments [3], there is a notable correlation between linguistic phenomena related to time and the kinds of mistakes made by both students and experts in translating English to LTL. This strongly suggests that the time-related natural language phenomena of Section II interact with the formalization process, and that confusion over whether or not to formalize these specific phenomena is a contributing factor to LTL misformalization. The results offer several takeaways for formal methods research. Direct education about these linguistic phenomena could help to reduce some of the most common kinds of LTL misformalization by humans, just as direct instruction about these phenomena aids in learning to communicate in natural languages [18], [31]. The fact that LLMs struggle with pragmatics and temporal expression in general [32]–[34] means LLM-based LTL formalization likely has similar weak points, which might either be avoidable by educated users or possible to improve with fine-tuning targeted to these specific weaknesses. Additional experiments are needed to investigate these and explore the impact of other factors in natural language expression of temporal ordering on LTL formalization.

ACKNOWLEDGMENT

This work was supported by US National Science Foundation Award #CCF-2220991.

REFERENCES

[1] C. Newcombe, T. Rath, F. Zhang, B. Munteanu, M. Brooker, and M. Deardeuff, "How Amazon web services uses formal methods," *Communications of the ACM*, vol. 58, no. 4, pp. 66–73, Mar. 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2699417

[2] M. Brooker and A. Desai, "Systems Correctness Practices at AWS: Leveraging Formal and Semi-formal Methods," *Queue*, vol. 22, no. 6, pp. Pages 60:79–Pages 60:96, Feb. 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3712057

[3] B. Greenman, S. Saarinen, T. Nelson, and S. Krishnamurthi, "Little Tricky Logic: Misconceptions in the Understanding of LTL," *The Art, Science, and Engineering of Programming*, vol. 7, no. 2, pp. 7:1–7:37, Oct. 2022, publisher: AOSA, Inc. [Online]. Available: https://programming-journal.org/2023/7/7/

[4] B. Greenman, S. Prasad, A. Di Stasio, S. Zhu, G. De Giacomo, S. Krishnamurthi, M. Montali, T. Nelson, and M. Zizyte, "Misconceptions in Finite-Trace and Infinite-Trace Linear Temporal Logic," in *Formal Methods*, A. Platzer, K. Y. Rozier, M. Pradella, and M. Rossi, Eds. Cham: Springer Nature Switzerland, 2024, pp. 579–599.

[5] A. Pnueli, "The temporal logic of programs," in *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*. IEEE, 1977, pp. 46–57.

[6] C. S. Gordon, "The Linguistics of Programming," in *Proceedings of the 2024 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, ser. Onward! '24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 162–182. [Online]. Available: https://dl.acm.org/doi/10.1145/3689492.3689806

[7] ——, "Mocking Temporal Logic," in *Proceedings of the 2024 ACM SIGPLAN International Symposium on SPLASH-E*, ser. SPLASH-E '24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 98–109. [Online]. Available: https://dl.acm.org/doi/10.1145/3689493.3689980

[8] E. b. Y. Huang, Ed., *The Oxford Handbook of Pragmatics*, ser. Oxford Handbooks. Oxford, New York: Oxford University Press, Mar. 2019.

[9] H. P. Grice, "Logic and Conversation," in *Speech Acts*. Brill, Dec. 1975, pp. 41–58, section: Speech Acts. [Online]. Available: https://brill.com/display/book/edcoll/9789004368811/BP000003.xml

[10] D. Wilson and D. Sperber, "Pragmatics and time," in *Relevance Theory*. John Benjamins, Mar. 1998, p. 1. [Online]. Available: https://www.jbe-platform.com/content/books/9789027285560-pbns.37.03wil

[11] L. Horn, "First things first: The pragmatics of "natural order"," *Intercultural Pragmatics*, vol. 16, no. 3, pp. 257–287, Jun. 2019, publisher: De Gruyter Mouton. [Online]. Available: https://www.degruyterbrill.com/document/doi/10.1515/ip-2019-0013/html

[12] J. Nerbonne, "Reference Time and Time in Narration," *Linguistics and Philosophy*, vol. 9, no. 1, pp. 83–95, 1986, publisher: Springer. [Online]. Available: https://www.jstor.org/stable/25001233

[13] B. Comrie, *Tense*, ser. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press, 1985.

[14] R. Declerck, *WHEN-Clauses and Temporal Structure*, Jan. 1997.

[15] B. Greenman, S. Saarinen, T. Nelson, and S. Krishnamurthi, "Accepted Artifact for Little Tricky Logic: Misconceptions in the Understanding of LTL." [Online]. Available: https://doi.org/10.5281/zenodo.6988909

[16] C. S. Gordon, "Data for linguistic theories coincide with misformalization in temporal logic," September 2025. [Online]. Available: https://doi.org/10.5281/zenodo.17047204

[17] D. Altshuler, *Events, States and Times: An essay on narrative discourse in English*. De Gruyter Open Poland, Nov. 2016. [Online]. Available: https://www.degruyterbrill.com/document/doi/10.1515/9783110485912/html

[18] K. Antoniou, "Multilingual Pragmatics: Implicature Comprehension in Adult L2 Learners and Multilingual Children 1," in *The Routledge Handbook of Second Language Acquisition and Pragmatics*. Routledge, 2019, num Pages: 16.

[19] L. E. Smith and C. L. Nelson, "World Englishes and Issues of Intelligibility," in *The Handbook of World Englishes*. John Wiley & Sons, Ltd, 2019, pp. 430–446.

[20] C. S. Gordon and S. Matskevich, "Trustworthy Formal Natural Language Specifications," in *Proceedings of the 2023 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, ser. Onward! 2023. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 50–70. [Online]. Available: https://dl.acm.org/doi/10.1145/3622758.3622890

[21] A. Brunello, A. Montanari, and M. Reynolds, "Synthesis of LTL Formulas from Natural Language Texts: State of the Art and Research Directions," in *26th International Symposium on Temporal Representation and Reasoning (TIME 2019)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), J. Gamper, S. Pinchinat, and G. Sciavicco, Eds., vol. 147. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, pp. 17:1–17:19, iSSN: 1868-8969. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2019/11375

[22] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, "What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution," in *2009 IEEE International Conference on Robotics and Automation*, May 2009, pp. 4163–4168, iSSN: 1050-4729.

[23] R. Nelken and N. Francez, "Automatic translation of natural language system specifications into temporal logic," in *Computer Aided Verification*, ser. Lecture Notes in Computer Science, R. Alur and T. A. Henzinger, Eds. Berlin, Heidelberg: Springer, 1996, pp. 360–371.

[24] K. M. Finney, "The Application of Software Metrics to the area of Formal Specification by," PhD Thesis, City University of London, 1998.

[25] R. J. Vinter, "Evaluating formal specifications : a cognitive approach," PhD Thesis, University of Hertfordshire, 1998, publisher: University of Hertfordshire. [Online]. Available: https://researchprofiles.herts.ac.uk/en/publications/evaluating-formal-specifications-a-cognitive-approach

[26] A. J. Ko, B. A. Myers, and D. H. Chau, "A Linguistic Analysis of How People Describe Software Problems," in *Visual Languages and Human-Centric Computing (VL/HCC'06)*, Sep. 2006, pp. 127–134, iSSN: 1943-6106. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1698774

[27] M. M. Imran, P. Chatterjee, and K. Damevski, "Shedding Light on Software Engineering-specific Metaphors and Idioms," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ser. ICSE '24. New York, NY, USA: Association for Computing Machinery, Apr. 2024, pp. 1–13. [Online]. Available: https://dl.acm.org/doi/10.1145/3597503.3639585

[28] D. Mendoza, C. Hahn, and C. Trippel, "Translating Natural Language to Temporal Logics with Large Language Models and Model Checkers," in *Proceedings of the 2024 Conference on Formal Methods in Computer Aided Design (FMCAD)*, Oct. 2024.

[29] M. Cosler, C. Hahn, D. Mendoza, F. Schmitt, and C. Trippel, "nl2spec: Interactively Translating Unstructured Natural Language to Temporal Logics with Large Language Models," in *Computer Aided Verification*, C. Enea and A. Lal, Eds. Cham: Springer Nature Switzerland, 2023, pp. 383–396.

[30] J. He, E. Bartocci, D. Ničković, H. Isakovic, and R. Grosu, "DeepSTL: from english requirements to signal temporal logic," in *Proceedings of the 44th International Conference on Software Engineering*, ser. ICSE '22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 610–622. [Online]. Available: https://dl.acm.org/doi/10.1145/3510003.3510171

[31] K. Antoniou, "Multilingual Pragmatics," in *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd, 2023, pp. 1–6.

[32] S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, and P. Bhattacharyya, "PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12 075–12 097. [Online]. Available: https://aclanthology.org/2024.findings-acl.719/

[33] B. Ma, Y. Li, W. Zhou, Z. Gong, Y. J. Liu, K. Jasinskaja, A. Friedrich, J. Hirschberg, F. Kreuter, and B. Plank, "Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8679–8696. [Online]. Available: https://aclanthology.org/2025.acl-long.425/

[34] Y. Qiu, Z. Zhao, Y. Ziser, A. Korhonen, E. Ponti, and S. Cohen, "Are Large Language Model Temporally Grounded?" in *Proceedings of the 2024 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 7064–7083. [Online]. Available: https://aclanthology.org/2024.naacl-long.391/