Unpacking Ambiguous Structure: A Dataset for Ambiguous Implicit Discourse Relations for English and Egyptian Arabic

Ahmed Ruby[1] **Sara Stymne**[1] **Christian Hardmeier**[2]

[1]Uppsala University, Department of Linguistics and Philology
[2]IT University of Copenhagen, Department of Computer Science
{ahmed.ruby, sara.stymne}@lingfil.uu.se, chrha@itu.dk

Abstract

In this paper, we present principles of constructing and resolving ambiguity in implicit discourse relations. Following these principles, we created a dataset in both English and Egyptian Arabic that controls for semantic disambiguation, enabling the investigation of prosodic features in future work. In these datasets, examples are two-part sentences with an implicit discourse relation that can be ambiguously read as either causal or concessive, paired with two different preceding context sentences forcing either the causal or the concessive reading. We also validated both datasets by humans and language models (LMs) to study whether context can help humans or LMs resolve ambiguities of implicit relations and identify the intended relation. As a result, this task posed no difficulty for humans, but proved challenging for BERT/CamelBERT and ELECTRA/AraELECTRA models.

1 Introduction

Coherence is essential for effective communication in written or spoken language (Adornetti, 2015), and discourse connectives play a crucial role in achieving it by helping readers or listeners to infer the intended discourse relation holding between two text spans (Asr and Demberg, 2020). Listeners generally have little difficulty recovering the intended meanings with implicit connectives which are inferred between two juxtaposed independent sentences. They evidence this by combining lexical cues, general reasoning, and prosodic cues to effectively identify the implicit discourse relation. When interpreting ambiguous implicit relations, prosodic cues can be used for disambiguation in spoken language (Tyler, 2014; Jasinskaja, 2009), while semantics is essential in both speech and writing, ensuring effective communication and understanding. Consider for instance the following examples:

(a) John is tall, *so* she will ask him out.

(b) John is tall, *but* she will ask him out.

(c) John is tall. She will ask him out.

The discourse relations in both (a) and (b) can be understood by listeners and readers because the connectives "*so*" and "*but*", respectively, explicitly indicate the discourse relation. Although the implicit discourse relation is ambiguous in (c), listeners might be able to infer it through prosody. However, it is still an open question whether specific prosodic cues are helpful for disambiguation in this case. Moreover, disambiguation can also be achieved in written and spoken language using semantic cues (e.g., additional context), such as adding different preceding context sentences that can enforce either the causal or the concessive reading. For instance, the preceding context for the casual and concessive reading can be:

(1) *She prefers tall men.* John is tall. She will ask him out.

(2) *She prefers short men.* John is tall. She will ask him out.

The additional context can influence the ambiguous structure, suggesting a likely interpretation in (1) that her preference for tall men implies a causal relation, while her preference for short men indicates a concessive interpretation in (2).

We observed that the ambiguous structure of implicit relations arises when the first argument (Arg1) does not provide specific details about the event being described and can be influenced by additional context information. However, for some ambiguous examples or structures, there is a clearly preferred reading even without any context, unless there is extremely strong evidence for a different reading. Consider for instance the following example (adapted from (Carston, 1993))

(a) Max fell. John pushed him.

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 126–144

The preferred reading for this example is that the pushing caused the falling. However, there is another possible reading where Max fell first and was later pushed by John, but it needs extremely strong evidence to force this interpretation. This means that it is hard to figure out if other aspects than semantics contribute to inferring the intended reading.

In order to explore how ambiguous implicit relations can be successfully resolved by the listener, we plan to conduct, in future work, a controlled experiment on the impact of prosody without being disambiguated by the semantic component. To support this, this study presents a small dataset of "truly" ambiguous examples for implicit discourse relations for both English and Egyptian Arabic, which cannot be resolved in the absence of any context, so that it enables a future investigation of prosodic features. We create a set of sentences with an implicit discourse relation that can be ambiguously read as either causal or concessive with two different preceding context sentences forcing either the causal or the concessive reading. The dataset is validated by humans who read these sentences and filled in the intended implicit discourse connective by choosing the most appropriate option from the provided list of connectives.

We were able to identify the ambiguous structure of implicit discourse relations and propose a new set of principles to construct ambiguity in implicit discourse relations. This process led to the creation of a small dataset for English and Egyptian Arabic that was validated by human participants. As far as we are aware, this is the first dataset that addresses ambiguous implicit discourse relations.

Since human participants were able to identify the intended implicit connectives in a set of examples, we investigate whether language models like BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) can also fill in the implicit connectives in the examples correctly, which is a challenging task, as context barely influenced the choice made by these models.

2 Related Work

2.1 Discourse relation datasets

Although discourse relations have been extensively studied over the last two decades, leading to elaborate taxonomies and inventories of varying scope and levels of abstraction, it is still challenging to provide a general definition for implicit discourse relations (Jasinskaja, 2009). However, there are some inferred relation types that are considered in Wolf and Gibson (2004); Miltsakaki et al. (2005); Prasad et al. (2008); Lavid and Hovy (2010) and annotated implicit relations were covered in the Penn Discourse Tree Bank 2.0 (PDTB 2.0) (Prasad et al., 2008), which is the most popular resource. Moreover, there are discourse-annotated corpora that cover implicit relations in multiple languages, such as TED Multilingual Discourse Bank, or TED-MDB, which contains transcribed TED talks in English, German, Russian, European Portuguese, Polish, and Turkish (Zeyrek et al., 2020), as well as in individual languages following the PDTB approach, such as the Hindi Discourse Relation Bank (Oza et al., 2009) and the Chinese Discourse Treebanks for Chinese. (Yuping et al., 2014; Long et al., 2020).

2.2 Discourse relations and ambiguity

Ambiguous structures can signal multiple potential interpretations of implicit discourse relations, and the intended relation can be inferred by context or by drawing on one's background assumption (Verhagen, 2000). Our study focuses on ambiguous implicit discourse relations, where a two-part sentence implies various potential relations, and must be inferred by context. Considering the distribution of discourse connectives in both PDTB and LADTB as reported by (Alsaif, 2012), the connectives 'but' and 'so' are commonly used in English and Arabic (Pitler et al., 2008; Alsaif, 2012). This observation has inspired the present study to explore the implicit relations that can be expressed by these particular connectives.

3 Ambiguity in inferring implicit relations

Each discourse relation involves two arguments, which are typically expressed as two clauses or phrases (Cabrio et al., 2013). Muskens (2000) argues that underspecified representations must be ambiguous. Drawing from this notion, we have shaped our own study's approach to examining the first argument with ambiguity in mind. The results of the validation confirm that if Argument 1 does not provide information that is relevant to inferring a specific discourse relation, it is not possible to make an inference about that relation unless there are underlying assumptions or presuppositions. In this case, it may be necessary to look for additional

information from context to infer the implicit discourse relation.

The meaning of Argument 1 can be shaped and influenced by context if it carries a neutral connotation, and Argument 2 gives additional information or detail based on the event influenced by context. Consider the example in Figure 1, where Argument 1 "John is tall" in both sentences is unspecified and needs to be interpreted in the context of the sentence to determine the intended information conveyed by Argument 2 "she will ask him out". In the first sentence, the context helped Argument 1 convey a positive meaning to infer that she has a preference for tall men, and because John fits this preference, she will ask him out, while in the second sentence, the context helped Argument 1 convey a negative meaning to infer that she does not have a preference for tall men, but she will still ask John out, even though he fits this preference.

While there is a lot of evidence that the context can disambiguate the discourse relation structure (Nowak and Michaelson, 2020; Lichao, 2010), we still do not have a thorough understanding of how ambiguity in implicit relations is structured, and how they can be interpreted only by context. In this regard, this study examines whether different preceding context influence whether the causal or concessive reading is elicited. A dataset was created and validated to investigate this question in two languages (English and Egyptian Arabic).

4 Constructing data for ambiguous implicit discourse relations

Creating a dataset for implicit discourse relations that involve ambiguity can be a challenging task. This is because the ambiguous structure is not linguistically defined in a way that influences meaning. Furthermore, inferring implicit discourse relations can be difficult, since it requires a nuanced understanding of language and discourse. Therefore, we aim to investigate this gap by identifying the ambiguous structure of implicit discourse relations and proposing a method to build a dataset for inferring relations by context.

4.1 Principles of constructing ambiguity in discourse relations

The initial validation findings, which are detailed in Section 4.4, reveal several principles that can be used when constructing ambiguity in implicit discourse relations, such as:

1. The discourse relation between sentences or Arg 1 and Arg 2 should be implicit, where:
   (a) *Arg 1* and *Arg 2* are not connected by any structural connective, such as "so", "but", "because", etc., but the connective can still be inferred.
   (b) *Arg 2* should not contain a lexical item e.g., "this" or "that" which implies a presupposition already established by *Arg 1*. This is because these anaphoric pronouns refer to the fact expressed by the first sentence, so the second sentence presents an evaluation of that fact, due to the lexical semantics of the verb that follows these pronouns. (Jasinskaja, 2009)
   (c) *Arg 2* should provide supplementary information or clarification for *Arg 1*.
2. Arg 1 should convey a neutral meaning[1], and be influenced by context. For instance, "The apple is red" can be influenced and shaped by context to be positive or negative.
3. The discourse relation can only be inferred by context.

4.2 Data and design

We create a set of contrastive sentences pairs that deliberately contain discourse relation ambiguities, with their preceding context, where both Arg1 and Arg2 are identical, with ”but” versus ”so”, depending on if the context makes Arg 2 expected or unexpected. As shown in Table 1.

Context	Target_sentence
The car is very cheap	It’s 100,000, __ I’ll buy it.
The car is very expensive	It’s 100,000, __ I’ll buy it.

Table 1: Paired example with ambiguous implicit relations with two different preceding context sentences forcing either the causal or the concessive reading.

As you can see in Table 1, the preceding context sentence guides the speaker to the intended meaning of the target sentence whether the discourse relation between adjacent sentence is a causal or concessive relation, and thus the speaker can fill in the connective in these sentences depending on the context.

[1]Arg 1 can also contain contronyms, which have opposite or contradictory meanings such as "crazy prices" can have opposite meanings depending on the context, it could refer to the low prices, or it could refer to the high prices. However, when testing this principle, we realized that it may be inferred by drawing on one’s background assumption.

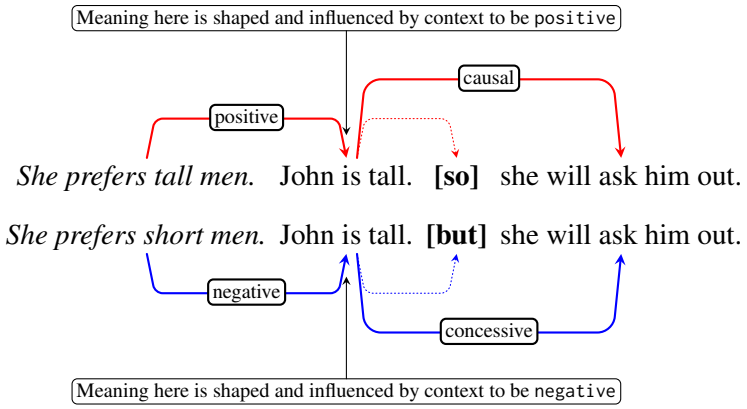


Figure 1: An example of inferring by context of the causal and concessive relation.

4.2.1 Arabic translation

There are five levels of Arabic used in Egypt as stated by Badawi (1973) in his socio-linguistic analysis of contemporary Arabic in Egypt: 1) Classical Arabic of the heritage, 2) Modern Standard Arabic, 3) Colloquial of the educated, 4) Colloquial of the enlightened, and 5) Colloquial of the illiterate. The "Colloquial of the educated" is a form of Arabic spoken by educated individuals that balance regional informality and linguistic proficiency. We opted for this level in our study, as opposed to Modern Standard Arabic (MSA) or "Colloquial of the enlightened," because it represents the prevalent form of spontaneous spoken communication. While MSA is widely understood, it is mainly used in formal written contexts or speeches, whereas "Colloquial of the enlightened" is characterized by its localized nature, which might limit understanding across regions or social groups.

By choosing the "Colloquial of the educated" level, we translated our English examples into Egyptian Arabic. In order to ensure consistency with the common writing style in Egyptian Arabic, two linguists, who are native Arabic-speaking, were asked to provide their feedback and suggestions about the writing style of the examples. This process helped to enhance the quality of the translation. After the first round of validation on translated examples, we decided to eliminate certain examples and introduce new ones. This implies that the English data is not entirely equivalent to the Arabic data.

4.3 Data validation method

In order to examine our data, we utilize human validation with the aim of ensuring the reliability and confidence of examples. This involves a number of procedures:

4.3.1 Distractors

To distract the respondents from the purpose of the study, and reinforce the impression that participants were reading the sentences naturally, we randomly interleave a number of distractors/ fillers with the target examples, which reflect the other implicit discourse relations: expansion and temporal according to the PDTB relations hierarchy (Prasad et al., 2008). Since distractors should be fitted syntactically in all examples, we created 5 examples with implicit 'in fact' connective for expansion relation, and also 5 examples with implicit 'when' connective for temporal relation. These distracted examples are similar to the target examples in terms of design and construction, where contain two discourse units, e.g. clauses or sentences, with proceeding context such as:

(a) *Writing on walls is illegal.* The teacher arrived early in the morning, _ we were painting on the wall.

(b) *Many people were thankful for the experience of traveling by car to Sharm El-Sheikh.* We tried to travel there by car, _ it was a very wonderful experience.

For Arabic, we used the same procedures as those applied in English, but we found that the equivalent of the "when" connective, *lámma*, can convey both causal and synchronous relations simultaneously, which means that this equivalent can be fitted in with both relations. As a result, we decided to eliminate it and use the "at the time/sāʕithā" connective instead for the temporal relation.

4.3.2 Questionnaire design

To achieve our aim of distraction, we added additional connectives, resulting in a balanced distribution of fillers and actual test connectives. As a result, the selection list comprises four connectives: "when", "so", "but", and "in fact", as illustrated in Figure 2.

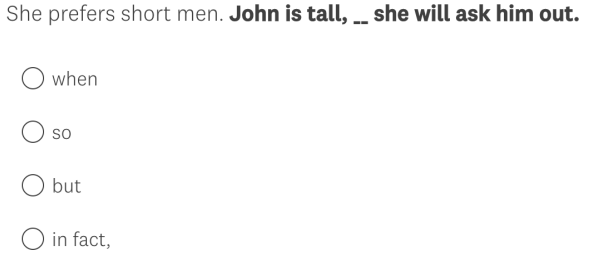
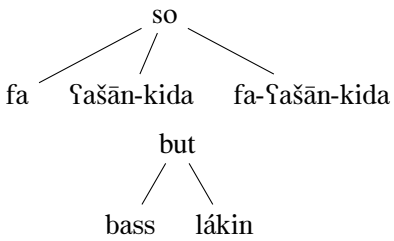


Figure 2: selection list

Regarding Arabic, there are more equivalent words for these connectives in Egyptian Arabic, such as "so" has three equivalents, and "but" has two equivalents as shown below.



For the "so" connective, (fa) the first equivalent can also be used as a filler, so we exclude it in this experiment, as it can be fitted in with other connectives, and the second and the third equivalent are similar in use, but the second is more widely used. We decided to use the second equivalent *ʕašān-kida* in the experiment.

The first equivalent of the "but/bass" connective is more commonly used, but the results of the first iteration showed that it can also be used as a filler, whereas the second one is mainly used in Modern Standard Arabic and among the educated in colloquial language. Therefore, we decided to use "lákin" as the equivalent of "but" connective in the second iteration.

We also found that the "in fact" connective has two equivalents, the first one is *fil-ħqīqa*, which can be also used as a filler, and the second one is *bil-fiʕl*, which is more commonly used among the educated in colloquial language. we ultimately decided to use the latter.

We use the same ratio of test items and fillers as that used in the English experiment in both validation iterations. In the first iteration, we used the Arabic equivalents of distractors/fillers used in the English experiment: "when/*lámma*", "so/*ʕašān-kida*", "but/*bass*" and "in fact/*fil-ħqīqa*". while in the second iteration, we replaced the "when" connective with "at the time/sāʕithā" and used another Arabic equivalent of "in fact" connective and "but" connective: "at the time/sāʕithā", "so/*ʕašān-kida*", "but/*lákin*" and "in fact/*bil-fiʕl*".

4.3.3 Participants

In order to investigate whether humans are able to identify implicit discourse connectives for these examples, we designed a questionnaire in English and invited volunteers with diverse native languages to answer the questionnaire by selecting the most appropriate connective from the provided list of options to fill in the blanks. In the first and second iterations, 24 and 21 participated in this validation, respectively.

We used the same process to create a questionnaire in Egyptian Arabic and invited Egyptian Arabic speakers to answer the questionnaire by selecting the most appropriate connective from the provided list of options to fill in the blanks as well. In the first and second iterations, 19 and 28 native speakers of Egyptian Arabic participated, respectively.

4.3.4 Procedure

To perform the validation process, we utilized the SurveyMonkey platform and enabled the randomization feature to randomize the two context sentences across participants so that each participant will only see one variant of each example. We distributed the survey link via an email list to gather responses from volunteers. In this task, we marked the main sentence in boldface, which contained the missing connective and preceded by context, and added a list of connectives under each sentence, as illustrated in Figure 2. The task was organized into three blocks of questions and was followed by a few language-related questions presented in Appendix C. However, these questions were not used for analysis. The entire validation task took approximately 10 minutes to complete.

The validation of the Egyptian Arabic dataset was also run by using SurveyMonkey. To collect responses from Egyptian people, we used Facebook to distribute the survey link and request their participation in answering the questions. Following the same processes of the English validation, where

Dataset	Validation	> 80% both	> 80% concessive	> 80% causal	< 80% both
English	1st iteration	1	3	13	14
	2nd iteration	19	3	8	1
Arabic	1st iteration	10	5	8	5
	2nd iteration	22	7	2	1

Table 2: The summary of human validation results on So/But groups of English and Egyptian Arabic datasets

each participant can only see one variant of each example and is not allowed to do the validation twice.

4.4 Results and Analysis

This section presents the summary of validation results on both English and Egyptian Arabic examples, showing the results of the So/But grouping that was conducted to compare the performance of the pair examples. Two validation iterations were conducted in both English and Egyptian Arabic. After analyzing problematic cases and refining our principles from the first English iteration, we created significantly improved examples for the second English iteration. Similarly, by examining the results of the first Egyptian Arabic iteration and adjusting the corresponding connective words in Egyptian Arabic, we achieved much better outcomes in the second Egyptian Arabic iteration. To ensure the validity and reliability of the validation process, a minimum threshold of 80% agreement between participants was established, meaning that only paired examples with a high level of agreement were included. Table 2 shows the summary of the validation results for both languages within each iteration. Detailed validation results can be found in AppendixA. We also employ Krippendorff's alpha to determine the degree of agreement or reliability among annotators/participants for each variant of an example. More detailed results of the inter-annotator agreement can be found in Appendix B.

In the first iteration of the English validation, which was performed on 31 paired examples, only one example in both cases: causal and concession met the threshold, while 14 examples in both cases did not meet the threshold. However, in the second iteration, the findings reveal that the participants were relatively successful in selecting the appropriate connective, with 19 examples in both cases meeting the threshold and only one example in both cases failing to meet the threshold.

In the first iteration of the Egyptian Arabic validation conducted on 28 paired examples, only 10 examples in both cases met the threshold, while 5 examples in both cases did not meet the threshold. However, in the second iteration, which involved 32 paired examples, the findings reveal that the participants were relatively successful in selecting the appropriate connective, with 22 examples in both cases meeting the threshold and only one example in both cases failing to meet the threshold.

To maintain consistency, we followed our principles and added three more examples to the existing set in English, so that both datasets contain a total of 22 examples.

For our preliminary dataset, the findings reveal that the participants were relatively successful in selecting the appropriate connective in the 'so' group. However, the results for the 'but' group were less promising, indicating that the participants struggled to identify the correct connective in these instances. This could be attributed to a range of factors, such as difficulties in understanding the intended meaning, or Arg 1 carries underlying assumptions or presuppositions. For instance, instead of considering the contextual cues in the examples below, several participants relied on their presuppositions about how to interpret the meaning of Arg 1. Here is the phrase "the weather changed" that can carry presuppositions as it can be to the better or to the worse:

(a) *One day it was nice and sunny so my family and I decided to go on a trip.* Suddenly the weather changed, **[...]** we decided not to go.

(b) *On the morning of the game, it was cloudy and rainy.* Suddenly the weather changed, **[...]** we decided not to go.

Here is also the phrase "The time was short" typically implies a negative outcome rather than a positive one, especially when the second argument indicates the result of the event:

(a) *The guest lecturer we had this week was much less long-winded than our usual professor.* The time was short, **[...]** I had fun.

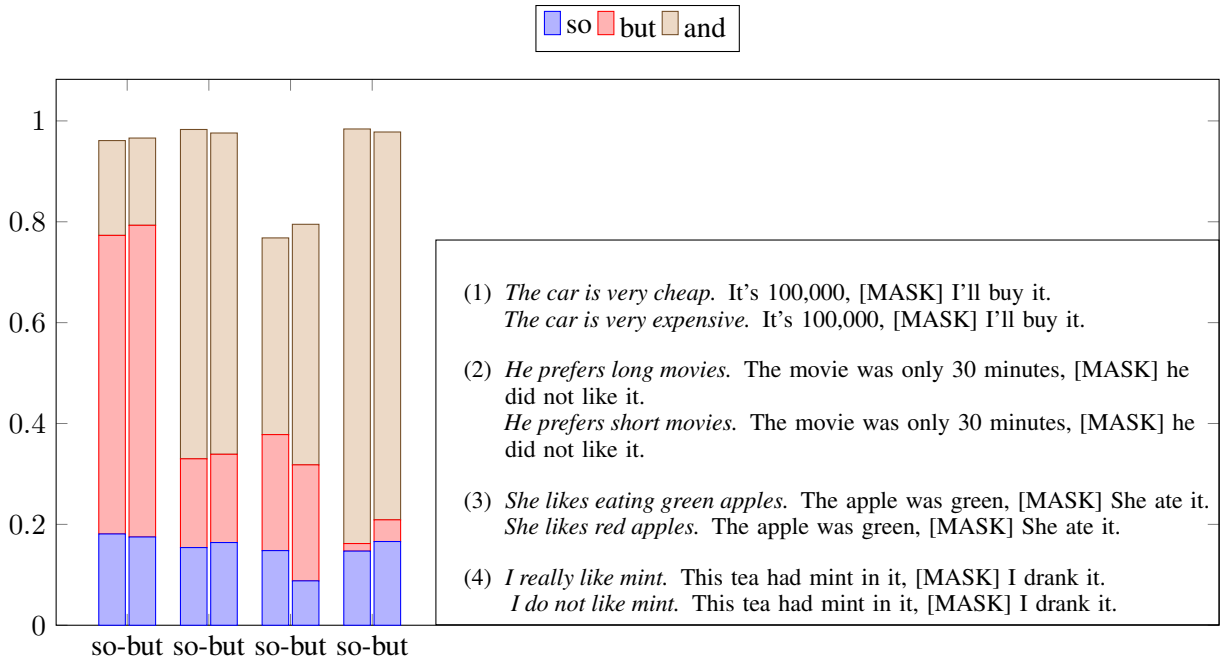


Figure 3: Sample of results from the pilot experiment showing four examples of So/But groups in English along with their respective scores

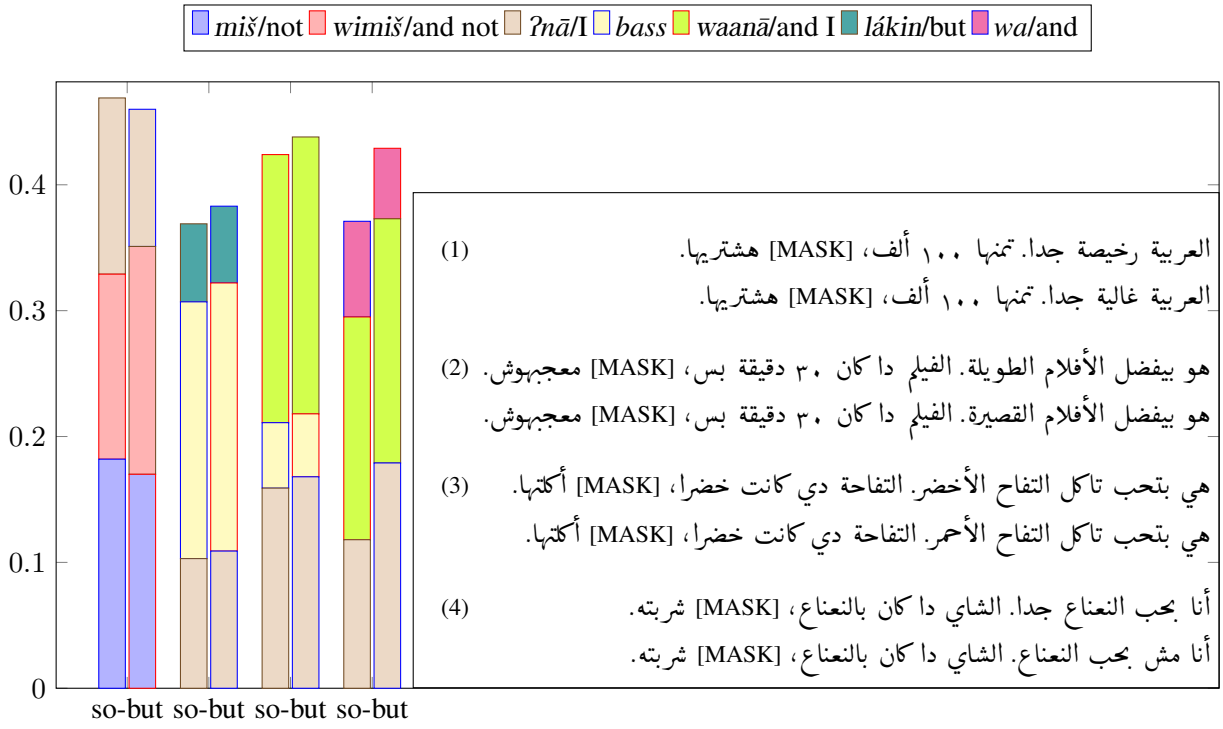


Figure 4: Sample of results from the pilot experiment showing four examples of So/But groups in Egyptian Arabic, which are the translations from English in the same order, and their respective scores

(b) *I spent a great time with my family.* The time was short, **[...]** I had fun.

Consequently, we removed instances with confusing or inconsistent results, modified some phrases to improve clarity, and added new instances. These changes led to a second iteration of human validation.

5 Pilot experiment

Since human participants were able to identify the intended implicit connectives in a set of examples, we now investigate whether language models like

BERT and ELECTRA will also be able to correctly fill in the implicit connectives within the provided examples.

5.1 English version

For English examples, we use the uncased version of the bert-base and electra-base models from Hugging Face[2] by inserting a mask between Arg 1 and Arg 2 to fill in the missing word, with setting up the topk parameter to 3 to obtain the top 3 predicted words.

5.2 Arabic version

We use CAMeLBERT-Mix (bert-base-arabic-camelbert-mix) model (Inoue et al., 2021) from Hugging Face as well, which is trained on a mixture of Modern Standard Arabic (MSA), Dialectal Arabic (DA) and classical Arabic (CA) variants, to fill in the implicit words for Arabic examples by inserting a mask between Arg 1 and Arg 2, with setting up the topk parameter to 3 to obtain the top 3 predicted words. we also use AraELECTRA-base-generator (Antoun et al., 2021) from Hugging Face, with the same setting.

5.3 Results and Analysis

The outcomes for the top three predicted words by BERT and ELECTRA on paired English examples are detailed in Appendix D. Figure 3 displays here the top predictions and their corresponding scores from BERT for 4 paired English examples with masked connectives.

These results indicate that identifying implicit discourse connectives is quite challenging for language models due to not capturing the influence of context on Arg 1 as there are small differences in the predictions for both So/But groups.

The results of the top three predicted words for Arabic examples, which encompass 25 and 33 different words in BERT and ELECTRA respectively, are also illustrated in Appendix D. Figure 4 presents here the top predictions and their corresponding scores from BERT for 4 paired Egyptian Arabic examples that include masked connectives. These examples are translations of the examples in Figure 3, following the same order.

These results indicate that identifying implicit discourse connectives for Arabic examples is quite challenging as well, as context barely influenced the choice made by these models. Furthermore, the performance of the models on Arabic examples is extremely poor, as many of the predicted words do not function as connectives. As shown in the legend entries of the figures, the words enclosed in **black squares** are connectives, while others are not. This can be interpreted for several reasons:

1. There are potentially systematic differences in the prevalence of implicit discourse relations in spoken data compared to written texts (Rehbein et al., 2016).

2. A discourse relation can be communicated by a pair of clauses conjoined by "and", but the sentences are not connected asyndetically(Jasinskaja, 2009; Rohde et al., 2018), For example, the Result relation can be communicated implicitly both with or without and, such as (Jasinskaja, 2009):

   (a) She fed him poisoned stew *and so* he died.
   (b) She fed him poisoned stew *and* he died.
   (c) She fed him poisoned stew. He died.

The connective "so" in (a) explicitly indicates a causal connection, but the same relation is successfully conveyed in (b) and (c), despite the absence of "so" or even "and".

This can explain the appearance of "and" in both the legend entries of English and Arabic results. Since “wa/and” is proclitic in Arabic, which is usually attached to the word (Habash, 2010), it may provide an explanation for the appearance of the words enclosed in **red squares** within legend entries of the Arabic results.

6 Conclusion and Future Work

In this paper, we introduced principles of constructing and inferring ambiguity in implicit discourse relations, and created a dataset for ambiguous implicit discourse relations, specifically causal and concessive relations for both English and Egyptian Arabic. We also validated both datasets by humans and language models (LMs) to study whether context can help humans or LMs resolve ambiguities of implicit relations and identify the intended relation. For future work, we plan to conduct a controlled experiment on the impact of prosody to figure out whether specific prosodic features correlate with the disambiguation of implicit discourse relations. We also intend to construct more examples to build a classification model to identify the two implicit discourse relations.

[2]https://huggingface.co/bert-base-uncased

Acknowledgements

We extend our thanks to Hannah Rohde for her valuable insights, and to Rima Haddad for her feedback on our Arabic texts.

References

Ines Adornetti. 2015. The phylogenetic foundations of discourse coherence: A pragmatic account of the evolution of language. *Biosemiotics*, 8:421–441.

Amal Alsaif. 2012. *Human and automatic annotation of discourse relations for Arabic*. Ph.D. thesis, University of Leeds, UK.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Fatemeh Torabi Asr and Vera Demberg. 2020. Interpretation of discourse connectives is probabilistic: Evidence from the study of but and although. *Discourse Processes*, 57(4):376–399.

El-Said Badawi. 1973. *Mustawayat al- arabiyya almu asira fi misr (Levels of Contemporary Arabic in Egypt)*. Dar almaarif-cairo, Cairo, Egypt.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *Computational Logic in Multi-Agent Systems*, pages 1–17, Berlin, Heidelberg. Springer Berlin Heidelberg.

Robyn Carston. 1993. Conjunction, explanation and relevance. *Lingua*, 90(1):27–48.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nizar Y. Habash. 2010. *Introduction to Arabic natural language processing*, 1 edition, volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ekaterina Jasinskaja. 2009. *Pragmatics and Prosody of Implicit Discourse Relations: The Case of Restatement*. Ph.D. thesis, Universität Tübingen, Germany.

Julia Lavid and Eduard Hovy. 2010. Towards a science of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22:13–36.

Song Lichao. 2010. The role of context in discourse analysis. *Journal of Language Teaching and Research*, 1:876–879.

Wanqiu Long, Bonnie Webber, and Deyi Xiong. 2020. TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803, Online. Association for Computational Linguistics.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain.

Reinhard Muskens. 2000. Underspecified semantics. In Klaus von Heusinger and Urs Egli, editors, *Reference and Anaphoric Relations*, pages 311–338. Springer Netherlands, Dordrecht.

Ethan Nowak and Eliot Michaelson. 2020. Discourse and method. *Linguistics and Philosophy*, 43:119–138.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The Hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161, Suntec, Singapore. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International*

Conference on Language Resources and Evaluation (LREC'16), pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Joseph Tyler. 2014. Prosody and the interpretation of hierarchically ambiguous discourse. *Discourse Processes*, 51(8):656–687.

Arie Verhagen. 2000. Concession implies causality, though in some other space. In Elizabeth Couper-Kuhlen and Bernd Kortmann, editors, *Cause - Condition - Concession - Contrast*, pages 361–380. De Gruyter Mouton, Berlin, Boston.

Florian Wolf and Edward Gibson. 2004. Representing discourse coherence: A corpus-based analysis. COLING '04, pages 134–140, USA. Association for Computational Linguistics.

Zhou Yuping, Lu Jill, Zhang Jennifer, and Xue Nianwen. 2014. Chinese discourse treebank 0.5.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A Parallel Corpus Annotated in the PDTB Style. *Language Resources and Evaluation*, 54:587–613.

A Human validation details

Figure 5 shows the results of the first iteration of human validation on English examples. This figure consists of two vertically stacked plots, each with four lines representing different categories: "When" (green stars), "so" (red circles), "but" (blue squares), and "in fact" (yellow triangles). The x-axis corresponds to the number of paired examples, labeled 1 to 31, while the y-axis represents the degree of agreement in responses. Each example has plotted points for each category.

In the first plot, the red "so" line has the highest values overall, with many points above 50. The green "When" line has some points above 20, but the majority of its points are below 20 or at 0. The blue "but" line has a few points above 10, but most of its points are at 0. The yellow "in fact" line is mostly below 20, with some points reaching above 20 or 30. On the other hand, the second plot shows the blue "but" with the highest values, featuring several points above 40 and the majority above 20. The red "so" line has several points above 20, but it is mostly below 40. The green "When" line is mostly below 20, with some points reaching above 20 or 30. The yellow "in fact" line has a few points above 20, but most of its points are at or near 0.

The results of the second iteration of human validation on English examples are shown in Figure6.

In the first plot, the red "so" line has the highest values overall, with many points ranging from 80 to 100. The green "When" line remains at 0 for all data points. The blue "but" line has a few points above 10, but most of its points are at 0. The yellow "in fact" line is mostly below 20, with some points reaching above 20. On the other hand, the second plot shows the blue "but" line with the highest values, featuring many points ranging from 80 to 100. The red "so" line is mostly below 40. The green "When" line is mostly at 0. The yellow "in fact" line has a few points above 10, but most of its points are at 0.

There is a significant improvement in both the "so" and "but" groups. As a result, we decided to select the example pairs that scored above 80% and translate them into Egyptian Arabic for further validation.

Figure 7 shows the results of the first iteration of human validation on Arabic examples. In the first plot, the red "so" line has the highest values overall, with many points ranging from 80 to 100. The green "When" line has a few points, but most of its points are at 0. The blue "but" line has some points above 10, and the majority of its points are below 20. The yellow "in fact" line is mostly below 20, with a few points reaching above 20. On the other hand, the second plot shows the blue "but" line with the highest values overall, featuring many points ranging from 80 to 100. The red "so" line is mostly below 40. The green "When" line is mostly at 0. The yellow "in fact" line has a few points above 10, but most of its points are at 0.

The findings indicate that using some Arabic equivalents as fillers led to confusion, making it challenging for participants to identify the correct connective in these cases. Therefore, we tried to avoid using ambiguous equivalents and proposed alternative equivalents of the selection list. These changes also led to a second iteration of human validation.

The results of the second iteration of human validation on Egyptian Arabic examples are shown in Figure8, indicating a significant improvement in both the "so" and "but" groups. In the first plot, the

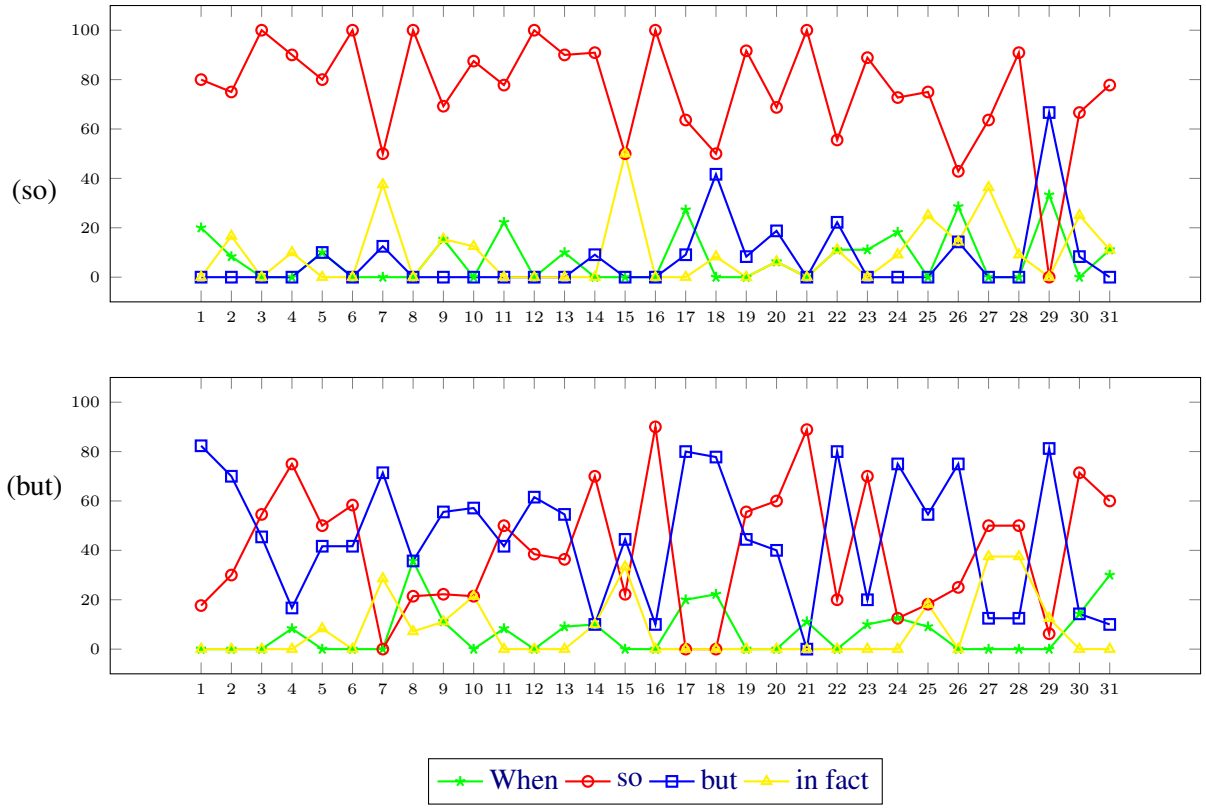


Figure 5: The validation results of the first iteration on So/But groups of English examples

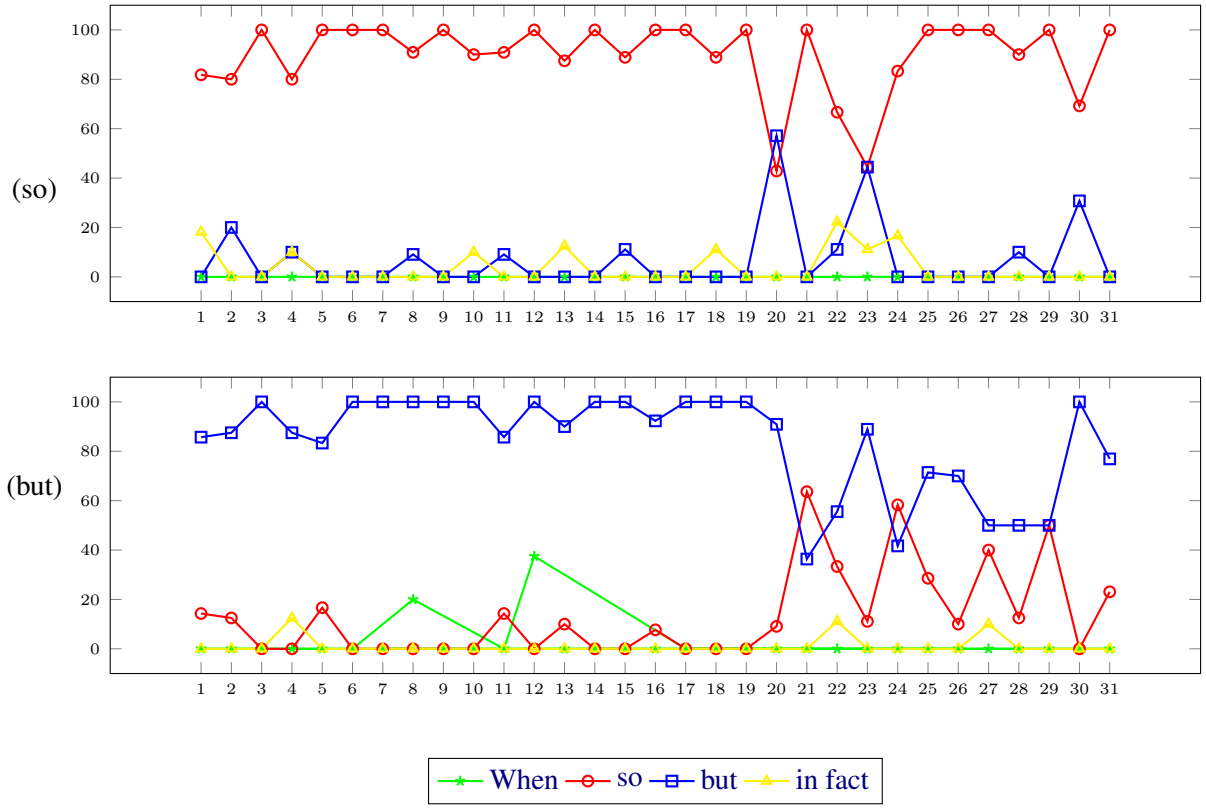


Figure 6: The validation results of the second iteration on So/But groups of English examples

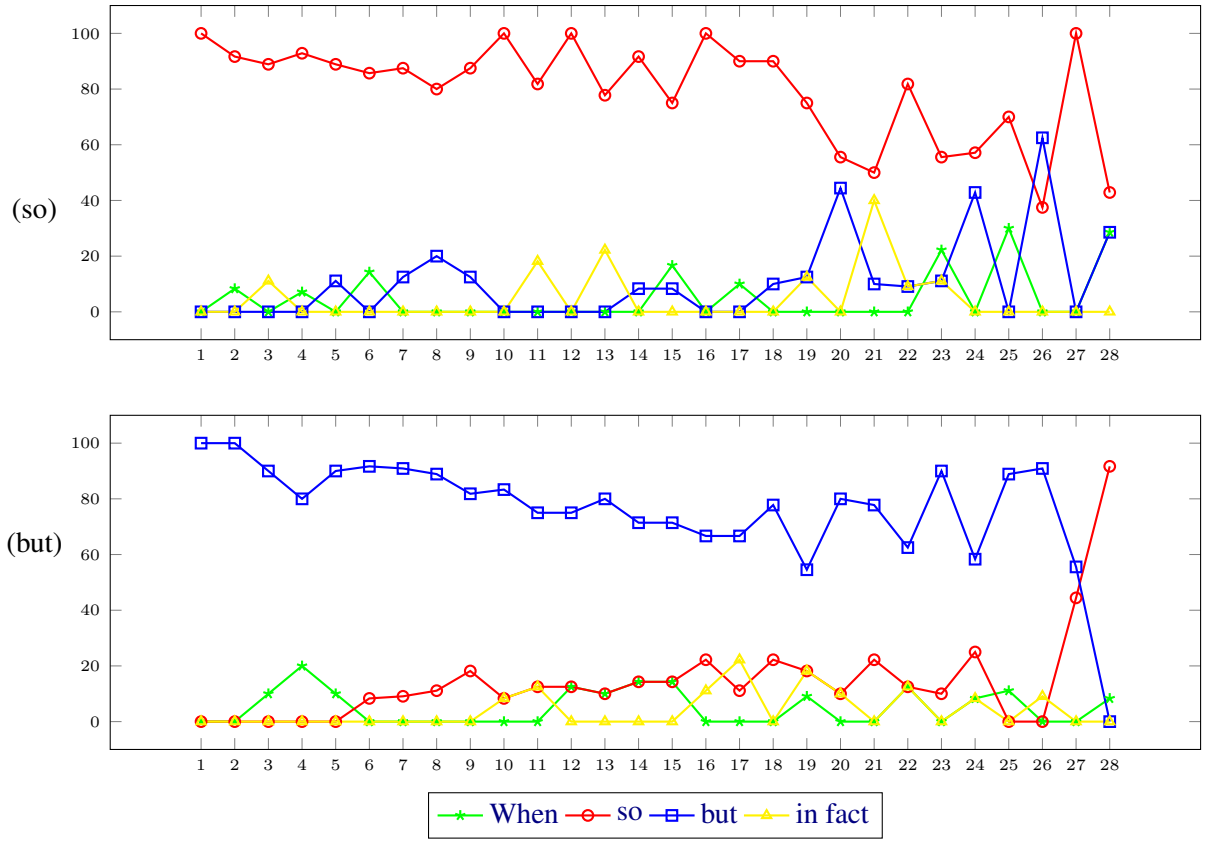


Figure 7: The validation results of the first iteration on So/But groups of Egyptian Arabic examples

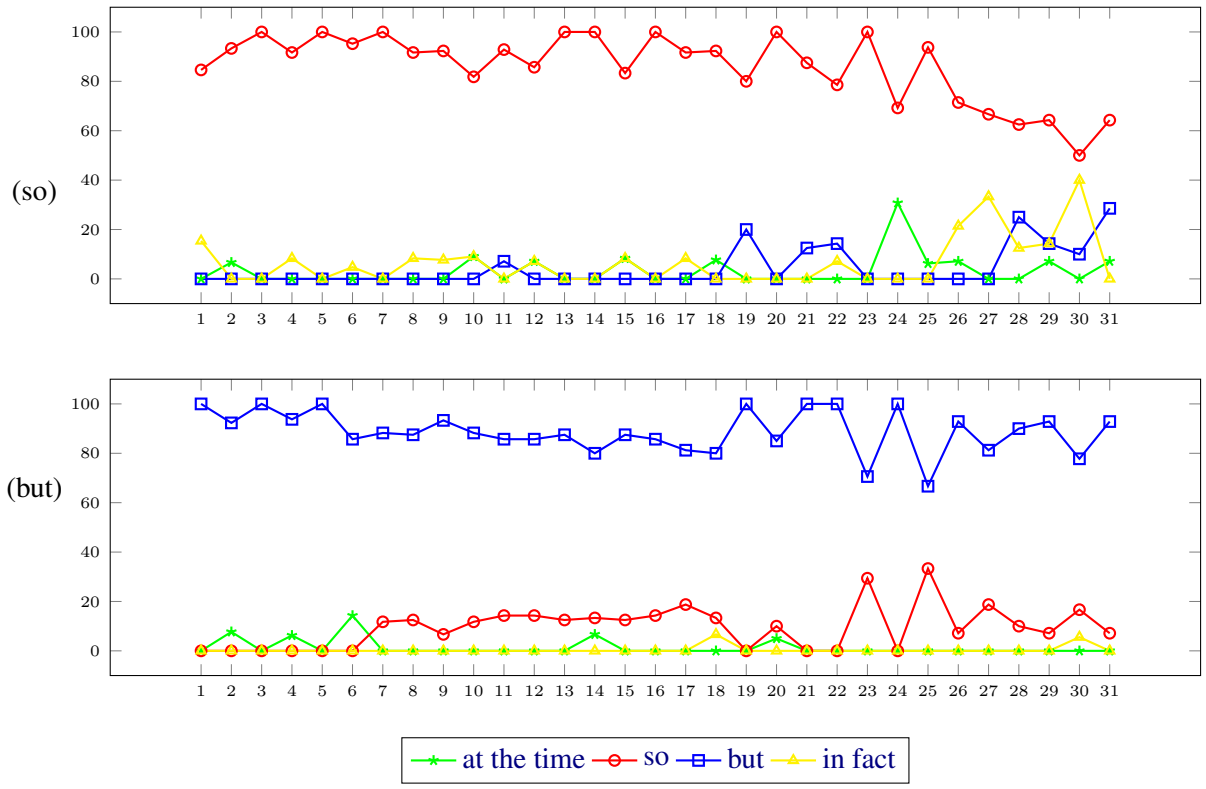


Figure 8: The validation results of the second iteration on So/But groups of Egyptian Arabic examples

Dataset	concessive	causal	concessive-causal pairs
English	22	22	22
Arabic	22	22	22

Table 3: Summary of the final examples for each discourse relation in each language

red "so" line has the highest values overall, with many points ranging from 80 to 100. The green "When" line has a few points, but most of its points are at 0. The blue "but" line has some points above 10, and the majority of its points are below 20. The yellow "in fact" line is mostly below 20, with a few points reaching above 20. On the other hand, the second plot shows the blue "but" line with the highest values overall, featuring many points above 40 and the majority above 20. The red "so" line has a few points above 20, but it is mostly below 40. The green "When" line has a few points above 10, but most of its points are at 0. The yellow "in fact" line has a few points above 10, but most of its points are at 0.

As a result, we obtained 19 (to which we later added 3 more, totaling 22) and 22 examples of pairs scoring above 80% for English and Egyptian Arabic, respectively. Table 3 provides a summary of the final examples count for each discourse relation in each language.

B Agreement Evaluation among Annotators

We use Krippendorff's alpha, which is a statistical measure to determine the degree of agreement or reliability among annotators/participants, by calling `krippendorff.alpha` function from the krippendorff Python package. Since each variant was only scored by a subset of all participants, we calculate it separately for each variant of each question, based only on the choice given by the subset of participants.

Table 8 shows the evaluation of inter-rater reliability using Krippendorff's Alpha calculation on the final English examples. It provides insights into the level of agreement among participants for each variant of an example, the concessive and causal relations. We observe that there are high agreement levels among the participants for most of the Causal and Concessive variants.

Table 9 illustrates the evaluation of inter-annotator agreement using Krippendorff's Alpha calculation on the final Egyptian Arabic dataset. The findings also reveal a substantial degree of agreement among the participants for the majority of the Causal and Concessive variants.

C Language-related questions

There were four language-related questions:

(1) *What was the first language you learned as an infant?* Table 4 displays a summary of the responses.

Dataset	Validation	en	sv-SE	ar-EG	other
English	1st iteration	6	7	0	7
	2nd iteration	7	5	0	12
Arabic	1st iteration	1	0	20	0
	2nd iteration	1	0	27	0

Table 4: The summary of answers for this question

(2) *Were any other languages spoken by your cares at home before you were 6?* Table 5 provides a summary of the responses.

Dataset	Validation	Yes	No
English	1st iteration	8	13
	2nd iteration	11	13
Arabic	1st iteration	0	21
	2nd iteration	0	28

Table 5: The summary of answers for this question

(3) *Did you attend daycare where a different language was spoken before the age of 6?* Table 6 shows a summary of the responses.

Dataset	Validation	Yes	No	en	ar-EG
English	1st iteration	4	17	1	0
	2nd iteration	6	14	4	0
Arabic	1st iteration	1	20	1	0
	2nd iteration	3	25	3	0

Table 6: The summary of answers for this question

(4) *What other languages do you speak fluently?* Table 7 shows a summary of the responses.

Dataset	Validation	No	en	fr	other
English	1st iteration	2	15	2	2
	2nd iteration	6	10	4	4
Arabic	1st iteration	9	12	0	0
	2nd iteration	13	15	0	0

Table 7: The summary of answers for this question

Dataset	Sentence Pair No	Causal		Concessive	
		No. of Participants	Agreement	No. of Participants	Agreement
English	1	11	0.57	10	0.74
	2	12	0.67	9	0.71
	3	6	1.00	15	1.00
	4	13	0.63	8	0.67
	5	6	1.00	15	0.67
	6	13	1.00	8	1.00
	7	12	1.00	9	0.71
	8	13	0.79	8	1.00
	9	13	0.79	8	1.00
	10	12	0.78	9	1.00
	11	13	0.79	8	0.67
	12	11	1.0	10	1.00
	13	11	0.76	10	0.74
	14	15	1.00	6	0.57
	15	6	1.00	15	0.55
	16	11	0.70	10	1.00
	17	7	1.00	14	0.81
	18	7	1.00	14	1.00
	19	11	0.76	10	1.00
	20	2	1.00	2	1.00
	21	2	1.00	2	1.00
	22	2	1.00	2	1.00

Table 8: Evaluation of Inter-Rater Reliability: Krippendorff's Alpha Calculation on the Final English Dataset Using the Nominal Measurement Level

Dataset	Sentence Pair No	Causal		Concessive	
		No. of Participants	Agreement	No. of Participants	Agreement
Arabic	1	8	0.67	7	1.00
	2	8	0.67	7	1.00
	3	15	1.00	2	1.00
	4	13	0.79	2	1.00
	5	5	0.67	10	0.74
	6	8	1.00	7	1.00
	7	8	1.00	7	0.63
	8	7	0.63	8	0.67
	9	8	0.67	7	1.00
	10	8	0.67	7	1.00
	11	8	0.67	7	0.63
	12	9	0.71	6	0.57
	13	6	1.00	9	0.71
	14	7	1.00	8	0.67
	15	7	0.63	8	0.67
	16	6	1.00	9	0.71
	17	7	0.63	8	1.00
	18	8	0.67	7	0.63
	19	8	1.00	7	1.00
	20	4	1.00	11	0.54
	21	10	1.00	5	1.00
	22	3	1.00	3	1.00

Table 9: Evaluation of Inter-Rater Reliability: Krippendorff's Alpha Calculation on the Final Egyptian Arabic Dataset Using the Nominal Measurement Level

D BERT and ELECTRA Validation

Figure 9 below shows the results of the top 3 predicted words on paired examples that scored above 80% in both cases in human validation for English. The figure presents the results of BERT on grouped examples through a pair of vertically aligned stacked bar charts. Each group represents the top predictions for masked connectives and their scores, which are the same in both groups (So, But, and And). In the first bar chart, the values of "so," "but," and "and" are distributed across the 22 bars/examples, with some bars showing a higher proportion of "so" or "but," and others displaying a higher proportion of "and." The second bar chart exhibits a similar distribution pattern. This means that identifying implicit discourse connectives is quite challenging for language models due to not

capturing the influence of context on Arg 1 as there are no differences in the predictions for both So/But groups.

The results of the top 3 predicted words for Arabic examples are illustrated in Figure 9. The figure presents the outcomes of BERT on grouped examples in Egyptian Arabic, utilizing a pair of vertically aligned stacked bar charts. Each group signifies the top three predictions for masked connectives along with their respective scores, which differ between the two groups. These predictions cover a total of 25 categories, with only 8 of them recognized as connectives.

In both plots, the highest values occur in categories 1, 4, and 19. Category 1 has the maximum value of 0.544, followed by category 4 with 0.482, and category 19 with 0.337. The results indicate that the model's performance on Arabic examples is extremely poor since a considerable number of the predicted words do not function as connectives. In the Results and Analysis section, I presented some interpretations for these results.

Figure 10 shows the results of the top 3 predicted words, which are "so", "but", "and", "because" and "where", by ELECTRA on paired examples for English. The figure shows the results of ELECTRA on grouped examples through a pair of vertically aligned stacked bar charts as well. The plot reveals that the outcomes for ELECTRA didn't differ much from the results of BERT. This observation further confirms that this task remains a substantial challenge for ELECTRA as well, primarily due to its limitations in capturing context.

The results of the top 3 predicted words for Arabic examples by AraELECTRA are illustrated in Figure 12. The figure shows the outcomes of AraELECTRA on grouped examples in Egyptian Arabic, utilizing a pair of vertically aligned stacked bar charts as well. These predictions cover a total of 33 categories, with only 11 of them recognized as connectives.

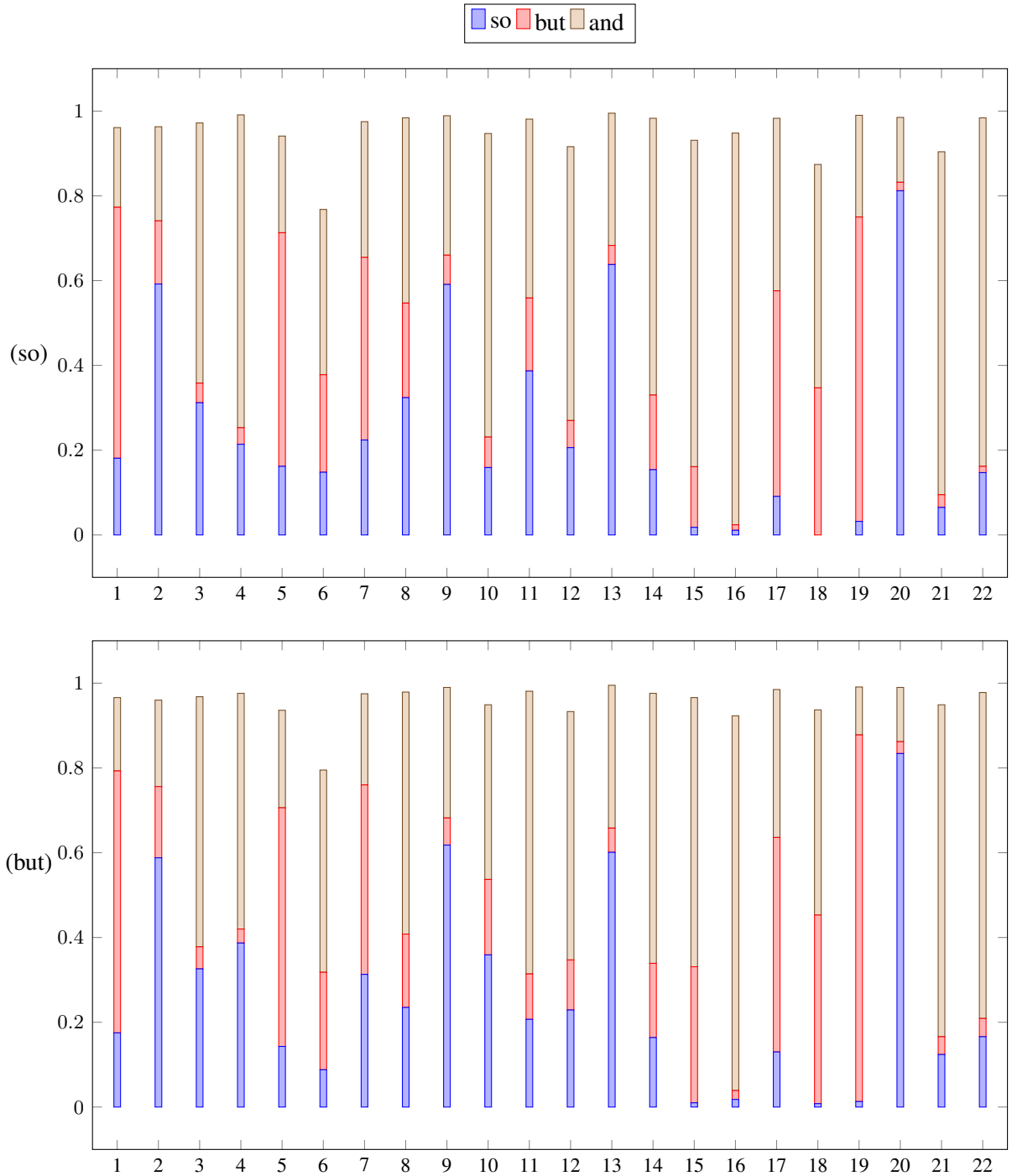


Figure 9: The validation results of BERT on So/But groups of English examples

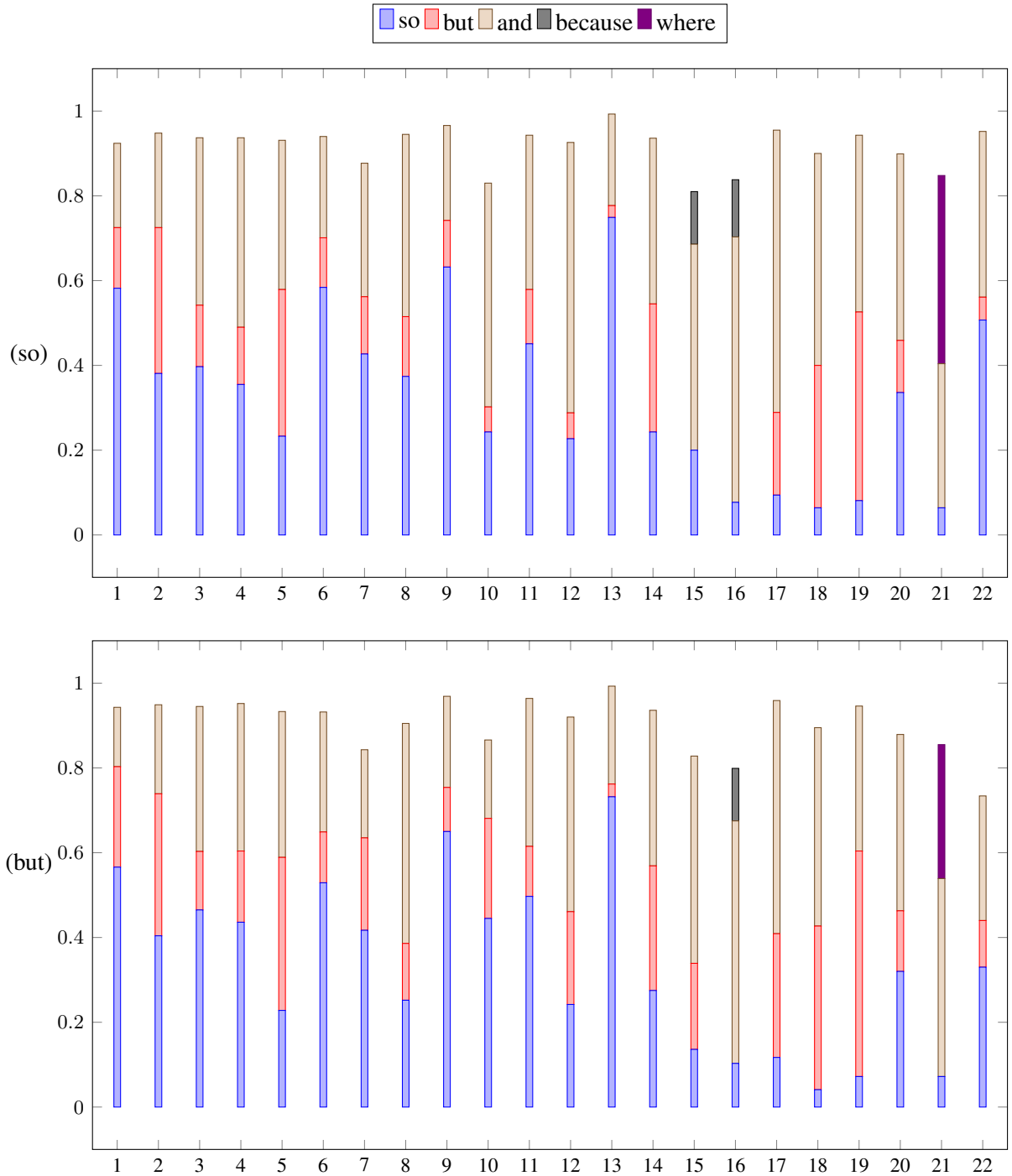


Figure 10: The validation results of ELECTRA on So/But groups of English examples

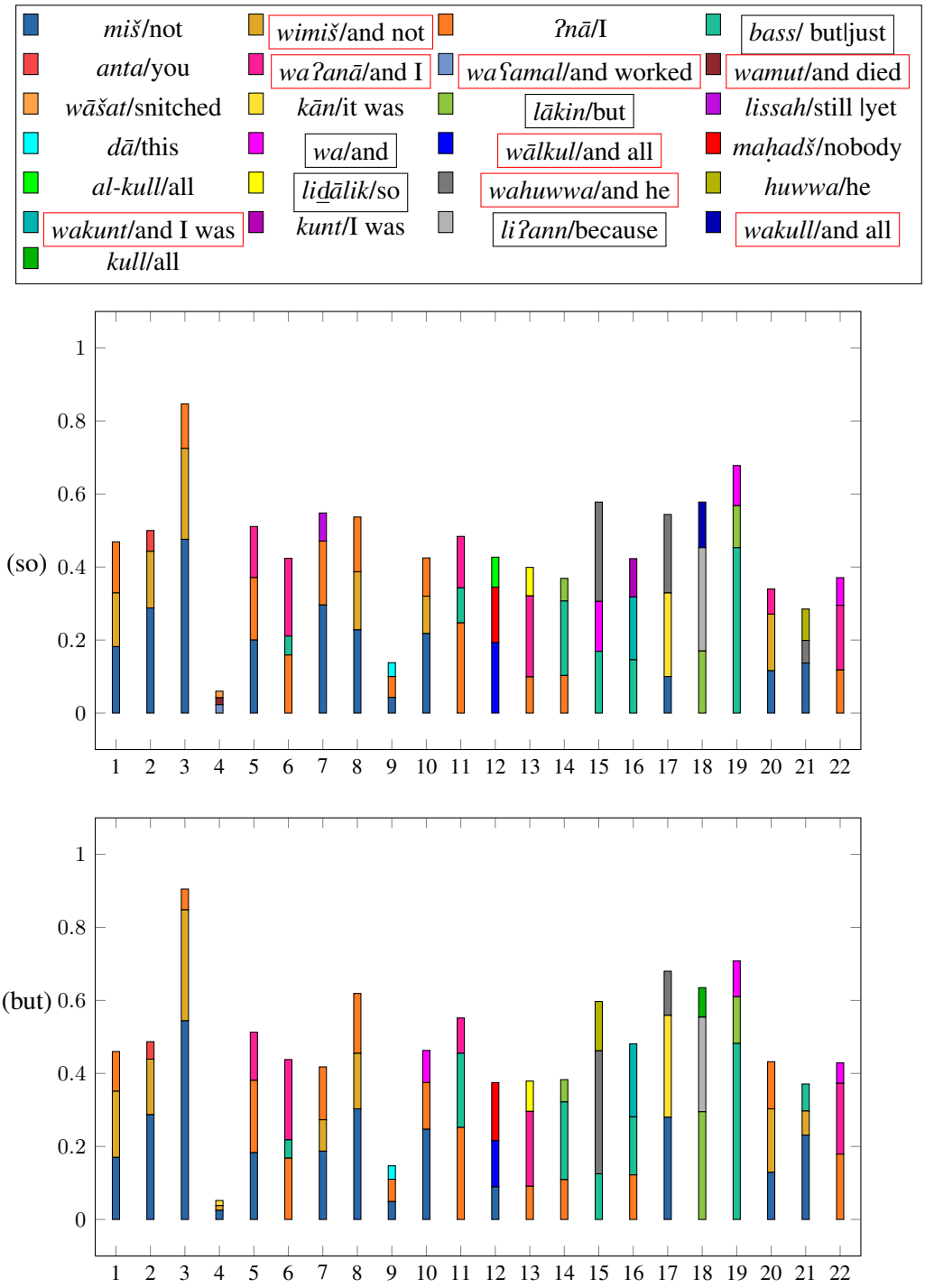


Figure 11: The validation results of BERT on So/But groups of Egyptian Arabic examples

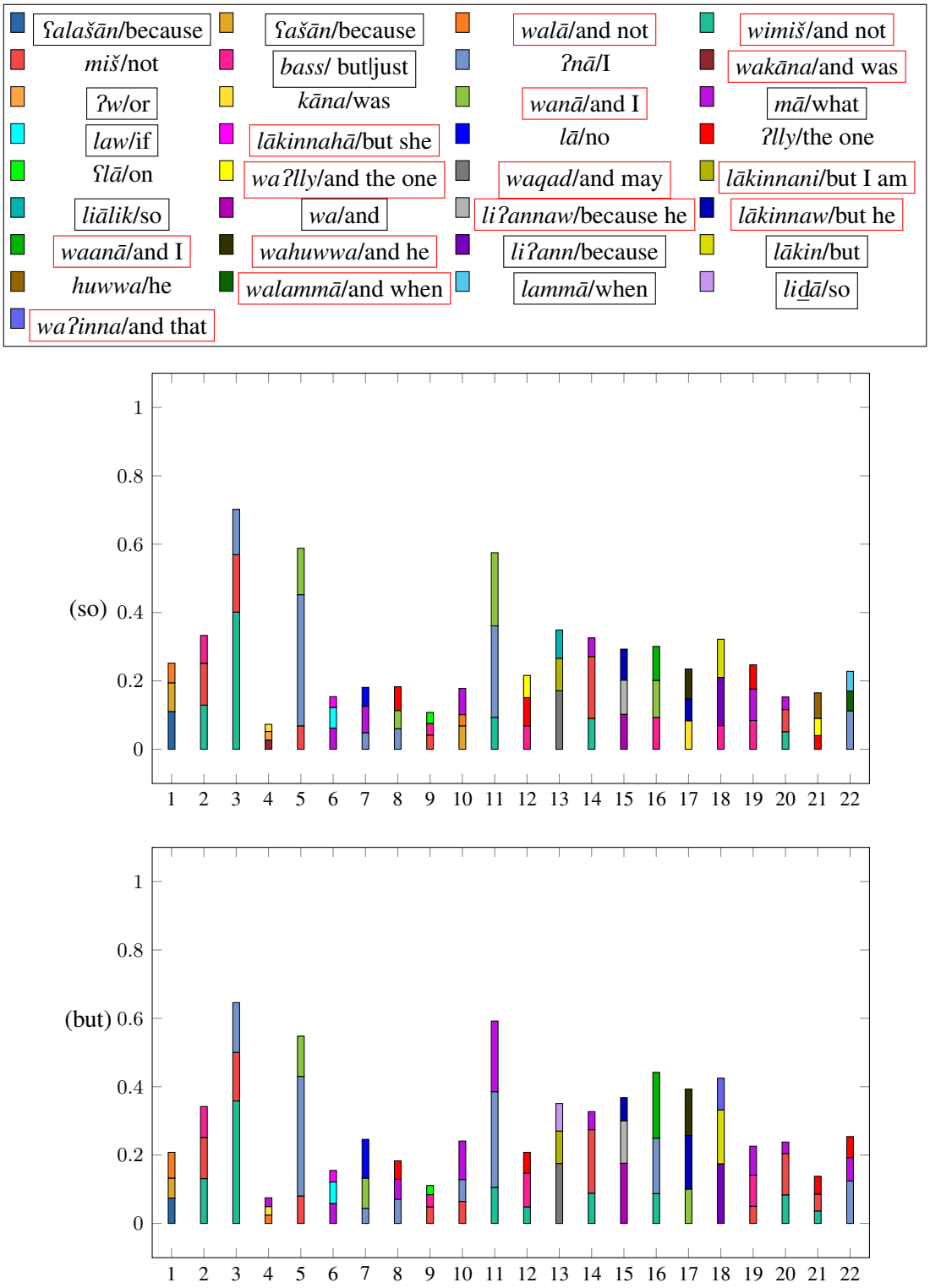


Figure 12: The validation results of AraELECTRA on So/But groups of Egyptian Arabic examples