Language Resources and Evaluation (2025) 59:3597–3633
https://doi.org/10.1007/s10579-025-09850-3

ORIGINAL PAPER

DiscoNaija: a discourse-annotated parallel Nigerian Pidgin-English corpus

Merel C. J. Scholman[1,2] · Marian Marchal[2] · AriaRay Brown[2] · Vera Demberg[2,3]

Accepted: 28 May 2025 / Published online: 20 June 2025

Abstract

This article presents a parallel English-Nigerian Pidgin corpus of PTB 3.0-style discourse relation annotations, named DiscoNaija. We explain the corpus design criteria, report inter-annotator agreement, and alignment and projection evaluations. We also present an update to a Nigerian Pidgin connective lexicon, named NaijaLex 2.0. An exploratory corpus analysis focused on comparing the distributions found in DiscoNaija to those found in PDTB 3.0 and a comparable corpus of English, DiscoSPICE. We identify various features of Nigerian Pidgin discourse coherence: (i) relations tend to be expressed implicitly more often in Nigerian Pidgin in general; (ii) anti-chronological temporal relations tend to be expressed less and are more likely to be expressed explicitly in Nigerian Pidgin; and (iii) coordinating conjunctions occur less frequently in Nigerian Pidgin than in English. The DiscoNaija corpus can facilitate a multitude of applications and research purposes, for example to function as training data to improve the performance of discourse relation parsers for Nigerian Pidgin, and to facilitate research into discourse features of creole languages.

Keywords Nigerian Pidgin · Discourse relations · Parallel corpus · Cross-linguistic comparison

✉ Merel C. J. Scholman
m.c.j.scholman@uu.nl

Marian Marchal
marchal@lst.uni-saarland.de

AriaRay Brown
ariaray@lst.uni-saarland.de

Vera Demberg
vera@lst.uni-saarland.de

[1] Institute for Language Sciences, Utrecht University, Trans 10, 3512JK Utrecht, the Netherlands

[2] Language Science and Technology, Saarland University, Campus C7.4, 66123 Saarbrücken, Saarland, Germany

[3] Computer Science, Saarland University, Campus C7.4, 66123 Saarbrücken, Saarland, Germany

1 Introduction

Nigerian Pidgin (also known as 'Naija') is an English-based contact language that developed as a result of European contact with West African languages. It is officially a pidgin, but it is widely considered to be an expanded pidgin or creole (Bakker, 2008; Faraclas, 2013; Parkvall, 2008). A creole language arises if a pidgin becomes the native and primary language of new generations of speakers. In the case of Nigerian Pidgin, there are over 100 million second language speakers, as well as 5 million native speakers (Faraclas, 2021).

Nigerian Pidgin and other pidgins and creole languages are characterized by unique features that make them interesting to study. They typically have a reduced vocabulary and simplified grammar, possibly making them more "efficient" languages, in the sense that they can convey similar concepts and information as more complex languages but at a lesser cost (Parkvall, 2008, p. 268). They also exhibit a rapid process of grammatical expansion and stabilization (Siegel, 2008), which makes them valuable for studying language evolution, acquisition, and grammaticalization processes (DeGraff, 1999). Additionally, their structures frequently blend features from multiple languages, offering insights into how different linguistic systems can converge and interact (Thomason, 2001). Pidgins and creoles can therefore provide valuable input for theories relating to the dynamics of language contact, the mechanisms of language creation, and how human cognition shapes language under unique social conditions.

In the field of linguistics, there is a marked imbalance in the focus on languages. English (and to a lesser extent a few other major languages such as Chinese, Spanish, and French) dominates linguistic research and computational applications (Blasi et al., 2022; Joshi et al., 2020). However, many languages remain underrepresented and under-resourced. This is also true for pidgins and creoles (Lent et al., 2022). As a result, there is a significant gap in both linguistic research and practical applications, limiting our understanding of the full range of human linguistic diversity and perpetuating inequalities in access to technology for speakers of underrepresented languages.

This focus on a limited number of languages is especially true for research on discourse structure and discourse relations (DRs) such as contrast and cause. Extensive discourse corpora exist for English (e.g., Carlson & Marcu, 2001; Webber et al., 2019), but are much smaller in other languages (Long et al., 2020; Zeyrek et al., 2020). Research on pidgins, in particular, is very limited. Yet pidgins and creoles present interesting opportunities for discourse analysis: their simplified clausal structures and more multifunctional, limited lexicon (Parkvall, 2008) may prompt speakers to rely on alternative strategies for coherence. For example, speakers might use fewer explicit connectives like *therefore* or *nonetheless*, and instead draw on non-connective cues or syntactic constructions that carry discourse functions. Pidgins and creoles also offer insights into the grammaticalization of connectives over

time (see Zufferey & Degand, 2024, ch. 5).[1] These features make Nigerian Pidgin a valuable testing ground for exploring discourse coherence and compensatory strategies in low-resource settings.

At the same time, Nigerian Pidgin poses real challenges for natural language processing applications: its low-resource status, limited morphology, and flexible syntax are conditions under which neural models often perform poorly (Ahmad et al., 2019; Ponti et al., 2020; Ruder et al., 2019). Expanding resources for Pidgin can therefore support the development of more robust and generalizable discourse parsers. These, in turn, can benefit downstream applications that depend on discourse understanding, such as machine translation, question answering, and coreference resolution. For instance, a discourse parser trained on Pidgin could enhance reference resolution in dialogue systems by identifying implicit temporal and causal relations between utterances.

A first step in investigating discourse marking in Nigerian Pidgin was undertaken by Marchal et al. (2021), who used a semi-automatic annotation approach to create a discourse connective lexicon of Nigerian Pidgin with English translations, called NaijaLex.[2] The current study builds on this work by presenting a discourse-annotated layer for an existing corpus of spoken transcribed Pidgin speech, named the Naija Treebank (Caron et al., 2019). The corpus covers a diverse range of topics, including life stories, speeches, radio programs, free conversations, cooking recipes, and comments on current states of affairs. The Naija Treebank is a parallel corpus, meaning that it consists of Nigerian Pidgin texts and their English translations.

The discourse-annotated version of the Naija Treebank that we present here, named **DiscoNaija**, includes annotations of explicit and inter-sentential implicit discourse relations in the PDTB3 framework (Webber et al., 2019). Specifically, the corpus contains 11,344 discourse relation annotations over a total of 140,859 words. The annotations are available for both the English texts and the Nigerian Pidgin texts. This resource can facilitate further research into discourse features of Nigerian Pidgin, as well as translation studies.

The main contributions of the present research are the following:

- We present a freely available parallel Nigerian Pidgin-English discourse-annotated corpus, DiscoNaija.[3] This is an annotation layer to the Naija Treebank, and consists of discourse relation annotations in PDTB3-style.
- We present an updated version of a Nigerian Pidgin connective lexicon, NaijaLex 2.0. This lexicon contains connectives, their translation equivalent, the PDTB 3.0 labels that they can express, and the frequencies in the DiscoNaija corpus.
- We study the feasibility of annotation projection from English to Nigerian Pidgin using a heuristic search strategy and NPMI to improve accuracy. The results

[1] For example, English *because* originated from the phrase *by cause*, meaning "by reason of". Over time, *by cause* was contracted and eventually fused into the conjunction *because*.

[2] See connective-lex.info for the implementation of this lexicon in a multi-lingual web app.

[3] https://osf.io/8m5vk/.

show that we can reach high accuracy in the projection of relation types and senses, as well as the alignment of arguments and connectives between English and Nigerian Pidgin.
- We explore unique features of discourse marking and structure in Nigerian Pidgin by comparing their distributions to those found in a comparable spoken discourseannotated corpus of English, DiscoSPICE. We take an exploratory approach to identify patterns in discourse organization across the two languages, thereby providing insight into how discourse relations are expressed in a low-resource contact language.

In what follows, we will first contextualize the current study by describing Nigerian Pidgin (Sect. 2) and reviewing related work (Sect. 3). In Sect. 4, we then describe the data included in the corpus and the annotation and projection procedures. Section 5 presents corpus statistics, including the distributions of relation senses and connectives, and a comparison between distributions of the DiscoNaija corpus and a comparable corpus of spoken English, DiscoSPICE. Section 6 discusses implications of these results and future directions.

2 Nigerian Pidgin

2.1 Origins

Nigeria is home to more than 500 languages. In this context, Nigerian Pidgin serves as a popular lingua franca. It developed in the seventeenth and eighteenth centuries as a simplified contact language during the Atlantic slave trade and British colonization (Faraclas, 2021). Due to the continuous and diverse interactions between multiple ethnic groups in Nigeria, Nigerian Pidgin expanded over the centuries, absorbing influences from various Nigerian languages. This has made Nigerian Pidgin more structurally complex and versatile, transforming it into what is often considered an expanded pidgin or a creole (Faraclas, 2013).

The Nigerian government does not officially recognize Nigerian Pidgin and there is no sanctioned role for the language in the education system, either as a medium or as a subject of instruction (Igboanusi, 2008). Schools mostly teach subjects using English as medium, which likely influences how Nigerian Pidgin speakers write in Pidgin: words are typically spelled and written as pronounced according to the sound patterns of Nigerian Pidgin, using a Latin-based alphabet (Esizimetor, 2009; Lin et al., 2024; Mensah et al., 2021; Ojarikre, 2013).

English is one of the main lexifiers of Nigerian Pidgin, with many words having a similar form and meaning as the English origin, as can be seen in Example (1).[4] Nigerian Pidgin is also influenced by other European languages, most notably Portuguese (illustrated by *pikin* in (1), stemming from the Portuguese 'pequenõ'), due to earlier Portuguese presence on the West African coast. Indigenous languages such as

[4] All examples of Nigerian Pidgin in this article are taken from the DiscoNaija corpus.

Yoruba, Hausa, and Igbo form the substrate of lexical, phonological, syntactic, and semantic influence on Nigerian Pidgin. Given this linguistic background of Nigerian Pidgin, we expect the majority of connectives to stem from the English lexicon.

(1) So I go close on time, go carry pikin go house.
So I finish on time, go get my child, and head home.

Nigerian Pidgin is a largely analytic language. This means that the verb is not conjugated, and that it uses auxiliary verbs before the verb to indicate tense or aspect. For example, to express that something has already happened or is completed, Nigerian Pidgin uses *don*, and to indicate a past event, it uses *bin* (see examples (2) and (3)) (Faraclas, 2004). This contrasts with English, which often marks tense by changing the form of the verb itself (e.g., English adds -ed to form the past tense in 'walked' or -ing for the progressive aspect in 'walking').[5] Because Nigerian Pidgin uses fewer of these morphological markers to show tense, speakers may rely more on temporal connectives like *before*, *after*, or *then* to convey the timing and sequence of events. This is in line with prior findings that adverbials and discourse connectives can compensate for a lack of verbal inflection (Bybee et al., 1994).

(2) When di man don see sey I don do di work finish, di man swallow di money.
When the man saw that I had done the work, he refused to pay up.
(3) I bin tell you! *I told you!*

Nigerian Pidgin follows a Subject-Verb-Object (SVO) word order, which means that in a typical sentence, the subject comes first, followed by the verb, and then the object, for example, "I eat rice." This basic sentence structure is similar to English, but the way Nigerian Pidgin connects actions within a sentence can be quite different. Like many West African languages, Nigerian Pidgin often uses serial verb constructions: a grammatical feature in which two or more verbs appear together in sequence to describe a chain of actions or closely linked events, all within a single clause. This is illustrated in Example (1), in which the second clause contains a serial verb construction (*go carry pikin go house*). In serial verb constructions, the verbs are not marked with prefixes or subordinating conjunctions to show that one depends on the other. They are also not typically linked by coordinating conjunctions like *and* or *but*, which are commonly used in English to join verbs or clauses (Aikhenvald & Dixon, 2005). This is also visible in Example (1): the serial verb construction in Pidgin requires the coordinating conjunction *and* to express the relationship between actions clearly when translated into English. Because Nigerian Pidgin uses serial verb constructions where English would typically use coordinating conjunctions, we expect to find fewer coordinating connectives overall in Nigerian Pidgin texts (Courtin et al., 2018).

[5] Note that English is also an analytical language, but less so than Nigerian Pidgin.

2.2 Linguistic resources

Despite Nigerian Pidgin being the most widely spoken pidgin/creole language in the world (Faraclas, 2021), linguistic resources on Nigerian Pidgin are limited. However, recent years have witnessed advances in the field of computational linguistics concerning creole and pidgin languages, driven by increased recognition of the significance of these languages and the need for diverse language resources (adapted to the needs of the community, cf. Lent et al., 2022). For example, Ogueji et al. (2021) trained a multilingual language model named AfriBERT on eleven African languages, including Nigerian Pidgin, using texts from the BBC. Similarly, existing corpora for Nigerian Pidgin (as well as Haitian Creole and Singaporean Colloquial English) were collected to release languagespecific language models by Lent et al. (2021). In an effort to enable NLP research on Creoles, Lent et al. (2024) introduced CreoleVal, a set of benchmarks covering a wide variety of tasks for up to 28 Creole languages, including Nigerian Pidgin. Lin et al. (2023) enriched existing available parallel and monolingual Pidgin datasets to generate a high-quality fully parallel corpus of Nigerian Pidgin text across ten resources and five domains. Lin et al. (2024) implemented a phonological-based word synthesizing framework to augment a Nigerian Pidgin dataset with orthographic variations, which improved performance on a sentiment analysis and a machine translation task.

Most important for our research is the Naija Treebank developed by Caron et al. (2019). As part of a larger project studying the syntactic and prosodic structure of Nigerian Pidgin, Caron et al. (2019) created a corpus of transcribed spoken data. The Naija Treebank contains Pidgin utterances (referred to as Source Text, ST), as well as their English translations (Translated Text, TT). The current study expanded this Naija Treebank corpus with a discourse annotation layer.

2.3 A Nigerian Pidgin connective lexicon

We build on previous work by Marchal et al. (2021), who used the Naija Treebank to create a lexicon of Nigerian Pidgin connectives. They exploited the English translations of the Nigerian Pidgin text and adopted a semi-automatic approach using automatic connective identification and annotation projection, combined with manual annotation. As a first step, they ran an automatic end-to-end PDTB classifier (Wang & Lan, 2015) on the English text to extract the TT connectives automatically and label them with PDTB2 relation senses. As a next step, Marchal et al. (2021) manually annotated the Nigerian Pidgin counterpart of a subset of English connectives to obtain a seed dictionary of English-Nigerian Pidgin connective mappings. This dictionary was then used to predict the Nigerian Pidgin equivalent of the English connectives in the remainder of the dataset. This approach led to the creation of NaijaLex, a Nigerian Pidgin connective lexicon complemented with automatic relation sense labels and the frequency data from the corpus. NaijaLex 1.0 consists of 57 unique connective types; the majority of which (n = 39) are derived from connectives in the English lexifier.

Based on an analysis of the automatically identified connectives, Marchal et al. (2021) concluded that Temporal and Cause were the most frequent explicit relations in the corpus. Conjunction relations were frequently implicit in Nigerian Pidgin texts and had been explicitated in the translations. These results provide tentative evidence for our expectation that temporal connectives would be relatively frequent in Nigerian Pidgin, and coordinating conjunctions such as *and* would be relatively infrequent in Nigerian Pidgin. However, the connective distributions do not allow us to draw definitive conclusions about relation frequencies, and the results from Marchal et al. (2021) also do not allow for comparisons with corpora in other languages and genres. To be able to draw more meaningful conclusions of discourse coherence in Nigerian Pidgin, we need discourse-annotated data that can provide insight into the frequency of relations (and their marking), rather than the frequency of connectives. This is the goal of the current study.

3 Discourse relation annotation and projection

Before turning to providing more details on the corpus creation method, we first present related work on discourse relation annotation and annotation projection.

3.1 Discourse relation annotation of (spoken) language

Discourse relations are semantic links between two arguments (Hobbs, 1979; Sanders et al., 1992; Webber et al., 2019). Following the Penn Discourse Treebank framework (Webber et al., 2019), we will refer to these arguments as Arg1 and Arg2. Relations that are marked by a discourse connective such as *because* or *however* are often referred to as explicit relations. Relations that do not contain a discourse connective are often referred to as implicit relations. For explicit relations, the argument that is syntactically bound to the connective is always labeled as Arg2 (cf. Webber et al., 2019); the other argument is Arg1. For implicit relations, the first textually occurring argument is always Arg1.

In order to study the distribution and linguistic realization of discourse relations, researchers use discourse-annotated corpora. One of the largest manually annotated discourse relation corpora available is the Penn Discourse Treebank (PDTB corpus, Webber et al., 2019). The framework that was used to annotate the corpus has also been used to create new corpora in other languages (e.g., Hindi, Oza et al., 2009; Chinese, Zhou & Xue, 2012; Turkish, Zeyrek & Er, 2022; Thai, Prasertsom et al., 2024) and genres (e.g., newspaper, Webber et al., 2019; TED talks, Zeyrek et al., 2020; biomedical texts, Prasad et al., 2011; novels, Scholman et al., 2022b; dialogues, Tonelli et al., 2010). This has resulted in different styles of PDTB annotation, but they can be considered to be interoperable (cf. Prasad et al., 2014). PDTB 3.0 is the most recent, state-of-the-art annotation framework. The DiscoNaija corpus that is introduced in this article is also annotated with PDTB 3.0 labels. We describe the premises of this framework in more detail in Sect. 4.

Most discourse-annotated corpora consist of written text, but since Nigerian Pidgin is primarily a spoken language, DiscoNaija contains transcribed spoken data. These two types of text are produced and processed differently (Chafe, 1982; Crible & Cuenca, 2017): spoken communication is characterised by a high degree of interactivity, sentence length tends to be shorter, and the pressure of rapid online processing often leads to disfluent structures. In contrast to written communication, the speaker and the hearer have access to additional channels of communication in spoken language, such as visual information or audio cues such as pitch and sentence stress. Although spoken language has access to these additional communicative channels, it also involves real-time planning and processing, which can increase the need for explicit discourse marking. Rather than relying on a wide range of specific connectives, spoken discourse tends to favor a smaller set of general connectives like *and*, *but*, and *so* (Rehbein et al., 2016). These connectives are highly frequent and multifunctional, serving a range of discourse functions depending on context (Schiffrin, 1987). They help structure discourse and guide listeners, particularly in the absence of the syntactic complexity or revision opportunities available in written language (Chafe, 1982). Note that Nigerian Pidgin is a primarily spoken language and that our corpus thus may also display a higher rate of explicitly marked relations compared to implicit relations.

Regardless of these differences between speech and writing, they do share a common set of relational meanings that can be annotated using the same relational framework (Crible & Zufferey, 2015; Rehbein et al., 2016; Tonelli et al., 2010), with some modifications to account for characteristics that are less likely to occur in written text, such as disfluencies. Although the PDTB framework was developed for written text, it has been applied to spoken data as well. Italian dialogues have been annotated in PDTB2-style in the LUNA corpus (Tonelli et al., 2010). PDTB3-style annotations have been used in annotations of TED talks in Chinese (TED-CDB, Long et al., 2020), a parallel corpus of TED talks in seven European languages (TED-MDB, Zeyrek et al., 2020), and a parallel corpus of implicit relations in novels and European Parliament proceedings (DiscoGeM, Yung et al., 2024). Finally, DiscoSPICE contains PDTB3style annotations on English spontaneous spoken data stemming from telephone dialogues and broadcast interviews (Rehbein et al., 2016). This dataset is most similar to the corpus we present here in terms of genre, and so we will use DiscoSPICE to compare discourse structure between Nigerian Pidgin and English.

3.2 Annotation projection and argument alignment

Many discourse-annotated corpora were created using a semi-automatic methodology, where automatic tools help identify or suggest discourse relations, but human annotators are crucial in refining and verifying the results to ensure that the discourse relations are linguistically accurate (e.g., Al-Saif & Markert, 2010; Prasertsom et al., 2024; Webber et al., 2019). One factor that impedes the creation of (parallel) discourse-annotated resources is that obtaining manual annotations is costly and time-consuming (but see Scholman et al., 2022c, for crowdsourced annotation

approaches). This challenge is especially pronounced for low-resource languages, where it is difficult to recruit annotators with both native-level fluency and the technical expertise required to apply complex discourse frameworks such as PDTB-style annotation. In the case of Nigerian Pidgin, while there is a large speaker base, there are very few trained discourse relation annotators with native fluency in the language, and no prior discourse resources or annotation guidelines tailored to its structure. This makes full manual annotation both logistically difficult and costly to scale.

Given these challenges, we adopt a projection-based approach. When dealing with a parallel corpus, it is far more efficient to annotate the higher-resource side (here, English) and then project these annotations to the lower-resource language. This strategy enables discourse-level annotation for languages with few trained annotators and no existing parsing infrastructure. While this comes with trade-offs, it allows us to produce large-scale resources where full manual annotation would not be feasible.

Indeed, projection approaches have been applied in prior discourse relation annotation efforts. However, most of this work focused on alignment or projection of explicit connectives, for example to create or extend connective lexicons and disambiguate connectives (Bourgonje et al., 2018; Das et al., 2020; Kurfalı et al., 2020; Laali & Kosseim, 2014; Mírovsky et al., 2021; Yung et al., 2023; Zhou & Xue, 2012), or to create a corpus annotated with connectives to train a discourse parser (Bourgonje & Lin, 2024; Laali & Kosseim, 2017; Versley, 2010). For example, Versley (2010) projected annotations of explicit English connectives, identified through an automatic discourse parser, to German text in an English–German parallel corpus. Similarly, Laali and Kosseim (2017) projected automatic annotations of English connectives to French connectives in European Parliament transcriptions. Our work differs in several ways. First, we project manually curated annotations, not automatic predictions. Second, we extend projection beyond explicit connectives to cover all full discourse relations, including implicit ones and their arguments. This provides insight into which relations are signaled and which tend to be expressed without an explicit connective. Finally, we target a spoken creole language for which no prior discourse-annotated resources exist.

Sluyter-Gäthje et al. (2020) were one of the first to apply a projection approach to the full discourse relation, as opposed to restricting the procedure to connectives. They created a German discourse-annotated corpus by automatically translating the English PDTB corpus and using word alignment to project the English annotations on the German target text. The current study takes a similar approach in that annotations from one language are projected onto another language using a parallel corpus. However, there are two differences in the types of text. First, the translations in the Naija Treebank are manual rather than machine translations. Second, the texts in the Naija Treebank are transcriptions of spoken dialogues and monologues, which contain more disfluencies. In addition, as described below, we use a slightly different algorithm to project connectives. Finally, we provide a more extensive evaluation of the quality of annotation projection by comparing the projected annotations with a manual gold annotation of the original text.

A common concern when projecting annotations from one language to another is that this approach relies on the assumption that coherence relations in the source

text and the translated text remain the same. However, this is not necessarily the case. Discourse relations often allow for multiple interpretations (Rohde et al., 2016; Scholman et al., 2022b), but a translated text might only contain the interpretation of the translator in cases where the original connective is disambiguated. Additionally, relations can be explicitated (a connective is added in the translated text) or implicitated (a connective is removed in the translated text). There are several factors that influence the implicitation or explicitation of a connective in translated text, such as specific features of the target language and the relation sense (Becher, 2011; Hoek et al., 2015, 2017; Lapshinova-Koltunski et al., 2022; Yung et al., 2023; Zeyrek et al., 2020; Zufferey & Cartoni, 2014). The distribution of relations in the translated text thus does not necessarily reflect the distribution of relations in the source text. This means that labels cannot simply be projected without taking into account the realization of the relation in the other language.

These concerns were validated in a study by Yung et al. (2024): they presented a manually annotated parallel corpus, and compared distributions of relations between the languages. The results revealed that the interpretation of implicit discourse relations does not always agree across the original texts and the translations, suggesting that discourse annotations might not always be projectable in parallel texts. However, this might in part be attributed to the fact that the annotations were performed by separate groups of annotators (e.g., one group of annotators per language) using slightly modified versions of the task. In the current study, we evaluate the projection accuracy by annotating the test set of the corpus in both languages. These annotations are done by the same annotators, thus allowing us to rule out inter-annotator disagreement effects.

3.3 Current study

We created a parallel discourse-annotated corpus of Nigerian Pidgin texts and their English translations. We did so by first annotating the English texts and then projecting the annotations to the Nigerian Pidgin texts. The corpus also allowed us to update the Nigerian Pidgin connective lexicon, presented here as NaijaLex 2.0.

A second goal of this contribution is to examine discourse structure and discourse marking in Nigerian Pidgin. We take an exploratory approach, using a coarse-grained analysis of English connective usage in Nigerian Pidgin and key features of the language to formulate expectations about how its discourse organization might differ from that observed in English. These expectations will inform our corpus exploration (Sect. 5).

First, relating to relation type distributions (e.g., whether relations are implicit or explicit), we expect Nigerian Pidgin to be characterized by a higher degree of explicit relations compared to other corpora. This is based on the observation that relations in spoken language are more likely to be marked with an explicit connective (Chafe, 1982; Crible & Cuenca, 2017; Rehbein et al., 2016). However, given that Nigerian Pidgin is characterized by less complex syntactic structures and a high degree of serial verb constructions (Courtin et al., 2018), it is also possible that relations in Nigerian Pidgin tend to be more implicit than relations in English spoken

language. We therefore compare the distributions in DiscoNaija to distributions of relations in PDTB 3.0 (English written language) and DiscoSPICE (English spoken language).

Second, we expect to replicate patterns of implicitness per relation sense that have been found to hold for English (Asr & Demberg, 2012). For example, condition relations tend to be marked explicitly, whereas level-of-detail relations tend to be marked implicitly relatively often. We will explore whether the ratio of implicitness for a particular relation sense diverges from that found in English data. Third, a particular relation for which we expect a difference between English and Pidgin data is the relation sense temporal asynchronous: we expect fewer implicit relations in Pidgin, because Nigerian Pidgin has fewer morphological markers for tense (Bybee et al., 1994; Faraclas, 2004).

Finally, relating to the connective lexicon, we expect that the majority of connectives will stem from English (cf. Marchal et al., 2021), and that coordinating connectives are less likely to occur in Nigerian Pidgin. This expectation is based on the greater degree of serial verb constructions that is typical for Nigerian Pidgin (Courtin et al., 2018).

4 DiscoNaija corpus creation methodology

This section describes the data that was annotated for DiscoNaija and how the data was annotated. Note that we first annotated the English translations of the corpus with discourse relations. We then projected these annotations onto the Pidgin portion. This section therefore also discusses how annotations were projected from English to Nigerian Pidgin.

4.1 Data

We added a layer of discourse annotations to an existing corpus of Nigerian Pidgin, namely the gold section of the Naija Treebank (UD NSC Corpus).[6] This is a parallel corpus of transcribed spoken Nigerian Pidgin utterances with English translations. The translation of the Nigerian Pidgin sentences into English was done by a team of native speakers of Nigerian Pidgin, and aimed at remaining as faithful as possible to the structure and style of the original utterances (Caron et al., 2019). The source data are spoken dialogues and monologues, but punctuation was added by the Naija Treebank annotators to reflect spoken rhythms and clause boundaries. The punctuation in UD corpora serves three main functions: it indicates sentence segmentation (periods, question marks, exclamations), it marks pauses or intonation breaks (commas), and it structures discourse (e.g., commas for discourse markers like *but*, *so*).

The dataset consists of 140,859 words (9242 utterances) collected in various locations across Nigeria. It is divided into three subsets: dev (n = 991 utterances),

[6] https://universaldependencies.org/treebanks/pcmnsc/index.html.

train (n = 7279), and test (n = 972). The data consists of recordings from 87 speakers. The sampling of speakers aimed at balancing age, sex, education, and linguistic and geographic background. The corpus covers a diverse range of topics, including life stories, speeches, radio programs, free conversations, cooking recipes, and comments on current states of affairs.

The original Naija Treebank corpus contains audio files with their transcription, utterances' translation into English, morphological tagging, macrosyntactic segmentation, dependency syntax, and prosodic annotation. These data can be merged with the DiscoNaija annotations for specific research purposes.

4.2 Annotation framework

The data were annotated using the PDTB3 framework. Discourse relations are taken to hold between two abstract object arguments, named Argument 1 (Arg1, presented in italics in examples) and Argument 2 (Arg2, presented in bold font in examples). In the DiscoNaija corpus, the arguments are utterances (as defined in the Naija Treebank, in the case of inter-sentential implicit or explicit relations) or parts of utterances (in the case of intra-sentential explicit relations). We adopted the utterance delineations from the original Naija Treebank. These delineations tended to be full sentences. All utterances are considered valid arguments, even if they, for example, consist of only a noun phrase (similar to the approach taken in Long et al., 2020).

4.2.1 Relation types

In addition to explicit and implicit relations, the PDTB distinguishes four additional label types. Alternative lexicalizations (AltLex) are alternative ways of lexicalizing discourse relations that lie beyond the closed set of discourse connectives, see (4) for an example (connectives and alternative lexicalizations are underlined in examples). This label was used when annotators inferred a relation between sentences but felt that the insertion of an implicit connective would be redundant.

(4) *Man fit no forgive.* Na why **de say "to err is human, to forgive is divine."** Man may not forgive. That's why they say, "to err is human, to forgive is divine."
[AltLex]

In Hypophora relations, one argument (commonly Arg1) expresses a question and the other argument (commonly Arg2) provides an answer, see (5).

(5) *How I wan take talk am o?* **Small ting no dey reach dem!**
How shall I put it? They are not satisfied with little things.
[Hypophora]

Entity relations (EntRel) represent identity relations between persons or objects mentioned in text segments, see (6). EntRel is annotated only when no semantic relation could be annotated between two adjacent text segments, but the utterances

did share the same entity.[7] We also annotated EntRel for relations where one argument consisted only of an interjection or similar type of words, like 'ah', 'mtschew', 'okay', 'hehehe', or 'wow' and both arguments were uttered by the same speaker, as in (7).

(6) *So di only way wey you fit take describe am na sey di animal na ojuju.* **So di ojuju dey catch people dat time.**
So, the only way you could describe it is that the animal is a monster. So, the monster was catching people at that time.
[EntRel]

(7) *Ehn!* **I know sey you go like am.**
Ehn! I know you'll like it.
[EntRel]

The NoRel label was used to annotate pairs of adjacent utterances that were neither related by a discourse relation nor by an entity relation, see (8).

(8) *As I con dey learn carpenter, I don dey sabi, I don dey sabi small, small.* **So de con call me for village say my moder no well.**
As I was learning carpentry, I was beginning to grasp, I was understanding little by little. So, they called me from the village saying that my mother was sick.
[NoRel]

As noted in Sect. 3, the PDTB framework is developed for written data. We added another label type to account for a feature that is specific to spontaneous spoken discourse: Interspeaker, see (9). This relation type was annotated when two adjacent utterances were spoken by two different people during a conversation.

(9) Speaker A: *Dream fit koba person.*
Speaker B: **Okay I don, I don hear you.**
Speaker A: Dreams can deceive you.
Speaker B: Okay, I've, I've heard you.
[Interspeaker]

4.2.2 Relational inventory

PDTB's relational inventory is structured as a three-level hierarchy, with four coarsegrained sense groups in the first level and more fine-grained senses for each of the next levels. The framework is presented in Table. 1. The top level, referred to as level 1, distinguishes four major semantic classes: Temporal, Contingency,

[7] Note that Arg2 in 6 contains so *so*, but in this example, it is used in a non-connective way.

Comparison, and Expansion. These classes are further refined in level 2. For example, the level Contingency contains relation labels for different Cause and Condition relation types, and Comparison is specified in different Contrast and Concession labels. The third level specifies the semantic contribution of each argument. For example, Concession has two subtypes: arg1-as-denier (Arg1 denies the expectation created by Arg2) and arg2-as-denier (Arg2 denies the expectation created by Arg1) (Table 1).

We added a feature (not a relational category) to the dataset to mark the completeness of the arguments. We used the tags *arg1-as-incomplete*, *arg2-as-incomplete*, or *both-incomplete* for relations where one of the arguments was interrupted. In some cases, one of these tags was annotated alongside a relation sense label, if the (assumed) intended relation could still be inferred. This is illustrated in (10), where it can be inferred that Arg2 is meant to convey that nothing besides breastmilk was given. When the annotators could not assign any meaning to the utterance, the relation was assigned a NoRel label and the tag marking the incompleteness (see (11)).

(10) *Because na only breastmilk I dey give am.* **I no give am any…**
Because I was only feeding her with breastmilk. I didn't give her any…
[Expansion.equivalence, *arg2-as-incomplete*]

(11) *If someone…* **If somebody…**
If someone… If somebody…
[NoRel, *both-incomplete*]

4.3 Annotation procedure

We first annotated the English translations of the corpus with discourse relations. We then projected these annotations onto the Pidgin portion (as will be addressed in Sect. 4.5). We followed PDTB's approach to relation annotation, which is a combination of manual and automated annotation: an automated process identified potential explicit connectives, and annotators then decided on whether the potential connective was indeed a true connective. If so, they specified one or more senses that held between its arguments. If no connective or alternative lexicalization was present (i.e., for implicit relations), annotators provided one or more connectives that together express the sense(s) they inferred.

To annotate explicit relations, we first identified potential explicit connectives. This was done using the PDTB e2e parser (Wang & Lan, 2015), which also provides PDTB2 relation labels, as well as a simple heuristic doing a string search of all PDTB3 connectives. Each candidate connective was manually inspected for connective status and annotated with a PDTB3 level-3 label, revising the automatically assigned PDTB2 label where necessary. Implicit relations were annotated by first inserting an implicit connective and then annotating this connective with a PDTB3 relation sense, as per PDTB guidelines. Annotators were encouraged to annotate multiple labels, especially for implicit relations. This better reflects the true meaning of the discourse relations, as relations can be ambiguous or even have multiple interpretations (Rohde et al., 2016; Scholman et al., 2022b).

Table 1 Relational inventory used to annotate DiscoNaija, based on PDTB 3.0

Top-level class	Type	Subtype
Temporal	Synchronous	
	asynchronous	precedence succession
Contingency	Cause	
	Reason	
	Result	
	negresult	
	cause + belief	reason + belief result + belief
	cause + speechact	reason + speechact result + speechact
	condition	arg1-as-cond arg2-as-cond condition + speechact
	negative-condition	arg1-as-negcond arg2-as-negcond negative-condition + speechact
	purpose	arg1-as-goal arg2-as-goal
Comparison	concession	arg1-as-denier arg2-as-denier
	concession + speechact	arg2-as-denier + speechact
	contrast	
	similarity	
Expansion	Conjunction	
	disjunction	
	equivalence	
	exception	arg1-as-excpt arg2-as-excpt
	instantiation	arg1-as-instance arg2-as-instance
	level-of-detail	arg1-as-detail arg2-as-detail
	manner	arg1-as-manner arg2-as-manner
	substitution	arg1-as-subst arg2-as-subst

4.4 Inter-annotator agreement

Annotations were done by three linguistically trained coders: the first three authors of this paper. These coders had trained together for other discourse annotation tasks. For the current task, they first trained on a subset of the data (550 utterances). This training consisted of four separate rounds of annotations, after which the coders

discussed any disagreements and necessary alterations to the relational inventory (see above). After training, a subset of the data was double-annotated by two coders to determine interannotator agreement (explicit relations: 5 texts, 398 connectives; implicit relations: 5 texts, 652 utterances). The remainder of the data was annotated by a single coder. Due to the ambiguous nature of implicit relations, all implicit relations were checked by another coder and disagreements were discussed. One coder focused only on annotation of the implicit relations, one only on explicit relations, and the other on both types.

Cohen's kappa (Cohen, 1960) is a metric frequently used to measure inter-annotator agreement (IAA). However, this measure is primarily used for comparison between single labels, whereas the annotations in DiscoNaija can consist of multiple labels. The traditional kappa is not suitable for evaluating the reliability of multi-label data, because it does note take into account that multi-label coding also inflates the chance agreement: by providing more labels, there is a higher chance that at least one of those labels overlaps with the annotations from another coder. We thus calculate inter-annotator agreement using a multi-label kappa metric (Marchal et al., 2022). This metric adjusts the multi-label agreements with bootstrapped expected agreement. Note that, when using this metric to measure agreement for data with single labels, it results in the same κ estimate as Cohen's κ.

For DR annotations, a $\kappa = .7$ is considered to reflect good IAA, whether it be Cohen's kappa or the multi-label kappa (Marchal et al., 2022; Spooren & Degand, 2010). Note that prior research has shown that agreement on implicit relations is more difficult to reach than on explicit relations, with a kappa score of .47 for PDTB level 3 senses in the Prague Dependency Treebank (Zikánová et al., 2019) and an F1 of .51 on crowdsourced annotations of implicits using a tagset with 7 level-2 labels (Kishimoto et al., 2018).

Table 2 presents the inter-annotator agreement on each of the levels of relation senses for explicit and implicit relations. Agreement ranges from sufficient to good for all sense levels except level 3 implicit relations, which is slightly lower than the desired kappa range, but is in line with prior literature on IAA for implicit relations.

4.5 Annotation projection

After completing annotation of the translated English texts (translated text, i.e., TT) in the corpus, we projected these annotations to the original Nigerian Pidgin text (the source text, i.e., ST). Compared to full manual annotation, projection allows faster corpus construction, while still yielding high-quality annotations with cross-linguistic interpretability (Laali & Kosseim, 2017; Meyer et al., 2011; Sluyter-Gäthje et al., 2020). Compared to fully automatic methods, projecting from human-annotated English source texts offers greater reliability, especially for complex discourse relations. Our approach thus provides a practical and scalable solution for bootstrapping discourse resources in under-resourced languages like Nigerian Pidgin.

Annotation projection relies on the assumption that discourse relations are preserved in translation. However, as discussed in Sect. 3.2, this assumption may not always hold: the discourse relation can be changed in the process of translation such

Table 2 Inter-annotator agreement for explicit and implicit discourse relations on the English portion of the corpus

Relation type	Sense level	%	Kappa
Explicit	Level 1	96	.94
	Level 2	85	.83
	Level 3	85	.83
Implicit	Level 1	84	.77
	Level 2	77	.72
	Level 3	64	.60

that the same overall content is expressed, but the discourse relation sense is different. This could be due to the discourse marking of the relations changing during translation: the connective may be translated as a more ambiguous variant, or the translated relation might not contain a connective although a connective was originally present (i.e. implicitated) (Crible et al., 2019; Yung et al., 2023). Note that this is less likely to occur in our dataset because Nigerian Pidgin tends to be more implicit than English. When a relation is underspecified or implicitated in translation, a sense shift can occur between the languages (Zeyrek et al., 2022): readers of the translated texts may not infer the same discourse relation as the readers of the original source texts. To assess the impact of these risks in our dataset, we annotated the test set of the Nigerian Pidgin texts (972 utterances of the 9,242 utterances in the corpus, annotated one year after annotating the English texts) and calculated intra-annotator agreement with annotations from the parallel English text.

In what follows, we first present the approach taken to project explicit and implicit relations, and then present the evaluation statistics.

4.5.1 Projection approach

The projection of the arguments of implicit relations was straightforward, since the corpus is utterance-aligned and the arguments of implicit relations consist of full utterances. The arguments of the TT explicit relations were aligned with AWESOME (Dou & Neubig, 2021), a neural word aligner that computes soft alignments based on word embedding similarity across languages. AWESOME is particularly useful in scenarios where traditional methods of alignment based on exact lexical matches may fail, especially when working with languages that differ in structure or vocabulary, like Nigerian Pidgin and English. AWESOME operates by leveraging multilingual word embeddings that capture the semantic similarity between words across languages. Instead of relying on direct translation pairs or word-for-word alignment, AWESOME computes a soft alignment. This means that it matches words from two languages (in this case, English and Nigerian Pidgin) based on their semantic proximity rather than their surface form. This approach allows AWESOME to handle cases where words in one language (like Nigerian Pidgin) may not have an exact equivalent in the other (like English), or where words in one language carry meaning that is distributed across multiple words in the other.

AWESOME relies on contextual embeddings to align words across sentence pairs. We use AWESOME without parallel finetuning, making it suitable for lower-resource settings where annotated parallel data may not be available. Specifically, we use the English tokenizer from SpaCy to segment and preprocess the input. While this model is not trained specifically on Nigerian Pidgin, its performance in aligning semantically similar words across English–Nigerian Pidgin pairs is sufficient for our projection tasks, particularly for identifying candidate connectives.

However, since AWESOME relies on semantic similarity, certain Pidgin words (such as interjections or reduplicated words) did not always align with specific English words, which led to discontinuous alignment spans. To address this, we implemented a postprocessing step: we manually ensured that skipped words (those not aligned to any English word) were included in the argument spans if they were not already accounted for in another argument or the connective. Additionally, when the connective itself was aligned to an argument (for instance, when a word in the connective matched with a word in the argument), we removed the connective from the argument span.

The projection of explicit relations was less straightforward, since these relations might occur within one or more utterances, and since this requires accurate mapping of the English connective to a Pidgin connective (if available). We evaluated several methods for projecting English explicit connectives onto the Nigerian Pidgin source text. The best-performing method, dict_project, relies on a seed dictionary and is described in detail below, as it is most relevant for understanding how the corpus was constructed. For settings where a seed dictionary is not available, we also tested a fully automatic method (awesome project). Finally, we experimented with a third method, awesome filtered, which combines the automatic approach with dictionary-based filtering.

4.5.1.1 Dict project The first approach we implemented to identify connectives in Nigerian Pidgin text is similar to that described in Marchal et al. (2021), but with some adaptations to account for PDTB 3.0's connective list and for more accurately aligning the best-fitting connectives.

We started by leveraging the NaijaLex connective lexicon. NaijaLex, however, was based on PDTB 2.0, which had a smaller set of connectives compared to the more recent PDTB 3.0. We updated the NaijaLex lexicon by incorporating additional English connectives from PDTB 3.0. We used this updated lexicon to run a heuristic search across the Nigerian Pidgin text. The goal of this search was to identify potential connective candidates in Nigerian Pidgin that could correspond to each of the English connectives annotated in the lexicon. To evaluate the quality of these connective candidates, we used the Normalized Pointwise Mutual Information (NPMI) metric. NPMI is a statistical measure that quantifies the strength of association between two words or phrases based on their co-occurrence in a corpus. In this case, NPMI was used to assess how well each potential Nigerian Pidgin connective candidate aligns with the English connective it is meant to correspond to. A higher NPMI score indicates a stronger semantic fit between the candidate and the intended discourse function of the English connective.

In the original methodology described by Marchal et al. (2021), the selection of connectives was based on NPMI values, with the relative position of the connective (i.e., its position within the sentence or clause) only being considered only when NPMI did not resolve to a one-to-one mapping. However, we made a modification to this approach by calculating a weighted score that incorporated both the NPMI and the relative position factors simultaneously. This adjustment was made because we found that, in practice, both semantic fit (as indicated by NPMI) and syntactic positioning are critical for identifying the correct connective in Nigerian Pidgin.

The original approach biases towards finding a connective, by assigning an implicit relation label only when no explicit connective was found. However, we wanted to avoid underestimating the rate of explicitation from source text (ST) to target text (TT), especially when there were explicit connective mappings in the TT English but implicit relations in the ST. To address this, we included explicit-to-implicit mappings in the analysis, giving them the same weight as explicit-to-explicit mappings in the scoring process. This ensures that both explicit and implicit discourse relations are taken into account when aligning the connectives across languages, reflecting the phenomenon where explicit connectives in English can correspond to implicit relations in Nigerian Pidgin.

In sum, when an explicit connective is annotated in English, dict project searches for a corresponding connective in the Nigerian Pidgin source text. If none is found, the relation is annotated as Implicit in Pidgin. For English Implicit relations, the Pidgin annotation is always kept implicit to avoid introducing false positives. This approach is applicable to any language pair with a seed dictionary (see Marchal et al., 2021, for guidance on constructing such dictionaries).

4.5.1.2 Awesome project For many languages, a seed dictionary may not exist. We therefore also implemented a fully automatic approach, inspired by Bourgonje and Lin (2024). We used AWESOME (Dou & Neubig, 2021) to align connectives and arguments. The annotations of connectives that were identified in English were then projected directly onto the aligned Pidgin word. However, an exploration of the development set of the dataset revealed that this approach sometimes selected Nigerian Pidgin words that are not connectives (e.g. *dere*, English: 'there').

4.5.1.3 Awesome filtered To address the issue of awesome_project aligning English connectives to Pidgin nonconnective words, we implemented a third approach, which is not fully automatic but rather a combination of the first and second approach. We filtered the projected Nigerian Pidgin connectives that were provided by awesome project such that this only contained words that occur in the NaijaLex connective lexicon (cf. Sluyter-Gäthje et al., 2020).

4.5.2 Evaluation of the projection approach

To test the performance of these three approaches, the first author annotated the connective (if any) and relation sense of each relation in the Pidgin test set.

Table 3 Performance accuracy of three different methods for projecting annotations

TT relation type	Dict project	Awesome project	Awesome project (Filtered)
Marking			
Explicit	0.95	0.89	0.88
Implicit	0.99	0.88	0.93
Combined	0.97	0.89	0.91
Connective			
Explicit	0.97	0.93	0.96
Implicit	–	0.05	0.08
Combined	0.97	0.81	0.88
Total			
Explicit	0.93	0.86	0.86
Implicit	0.99	0.89	0.93
Combined	0.96	0.87	0.9

Marking = agreement on whether relation is marked (related to implicitation/explicitation rate); *connective* = agreement on specific marker when a relation is marked; *total* = marking and connective together

We first evaluated the reliability of the relation marking projection—that is, the method's accuracy in estimating if a relation is marked and whether the appropriate ST connective was found. More specifically, we calculated accuracy for *marking identification* (i.e. agreement on whether the relation is marked), *connective identification* (i.e. given a projected ST explicit relation, agreement on whether the correct ST connective is found) and the *total agreement* (i.e. agreement on whether and how the relation is marked in the ST). Table 3 presents the accuracy for each of these measures.

Total agreement: Table 3 shows that dict proj most accurately determines whether and how the relation is marked, with average percentage agreement for identifying whether a relation holds and if the correct connective was aligned being 96%. Comparing the two remaining methods, awesome filtered consistently outperforms awesome project across most metrics. An exception is identifying whether a relation is marked in the source text, where awesome project has an advantage due to its ability to detect connectives not yet included in the dictionary. However, it also produces noisier results, occasionally retrieving non-connective terms like 'creche'. This highlights a trade-off: awesome_project is better for discovering new connectives, while awesome_filtered yields cleaner, more precise results. Additionally, both awesome project and awesome_filtered more frequently identify connective candidates that are not used in a discourse sense, as illustrated in (12). In this example, the TT connective *because* was added in translation, while the ST *sey* functions as a complementizer ('they know [that]'), not a discourse marker. Awesome project and awesome filtered, which do not consider word position, incorrectly align *sey* with *because*.

(12) Den go know sey we don do am finish.
Because they know we've finished.

Rate of explicitation: Of the 600 implicit Nigerian Pidgin ST relations, 8.8% (n = 54) are marked by a connective (excluding 1 AltLex) in the English TT. With respect to annotation projection from the English TT to the Nigerian Pidgin ST, this means that 10.6% of the explicit relations annotated in the TT will originally be implicit in the ST. For comparison, awesome_filtered marks 12% of explicit TT relations as implicit in the ST, which is relatively close to the gold rate of 10.6%.

When looking at some of the connectives that are missed in the alignment or projection, there is not a clear pattern that can be identified. The connective types that were missed more than once by all three methods are *and, also, like, when* and *still*. These are polysemous words that also function in non-connective usage (compared to, e.g., *because*, which is almost always a connective), which might explain why these were relatively more difficult to align.

Rate of implicitation: In Nigerian Pidgin to English translation, relations are less likely to be implicitated (i.e. a connective is removed) than to be explicitated (i.e. a connective is added). Of the 500 explicit Nigerian Pidgin ST relations, only 0.5% (n = 3) are left implicit in the English TT. The two alignment methods making use of AWESOME, awesome project and awesome filtered, allow for retrieving connectives from the Nigerian Pidgin source text.[8] However, the TT-ST explicitation rate is 11.3% for awesome project, or 6.7% when using awesome filtered. Retrieving ST connectives for implicit TT relations will thus yield a high number of false positives. For instance, the model often incorrectly retrieves 'na' as marking a relation when it actually is used as a focus particle ('it is me'), as in (13).

(13) Na me go forward am, go give oga.
I will be the one to forward it… to go and give it to my boss.

Intra-sentential explicit relations: Intra-sentential explicit ST relations that are implicit in the English TT are not covered by the dict_project approach. To examine what proportion of explicit Pidgin relations was found using our approach (i.e. recall of ST connectives), the first author additionally annotated all connective candidates in the test set (i.e. all words that occur in the connective lexicon, but are not yet annotated). Out of all ST connective candidates (n = 586) present in the test set, 82% were found using dict_project. Of the remaining connectives that were not found, for about one third no connective was present in English, either because the relation was implicitated, as in (14), or because there were no two separate clauses in English, as in (15). Note that the corpus only includes those intra-sentential implicit ST relations if they have been explicitated in the TT.

[8] Note that dict proj does not consider Nigerian Pidgin connectives for implicit relations, which is why no connective agreement is provided for in implicit relations.

Table 4 Distribution of relation types in DiscoNaija

Relation type	Count	Percentage (%)
Explicit	4952	43
Implicit	4952	43
AltLex	81	0.7
Hypophora	89	0.8
EntRel	592	5
NoRel	678	6
Interspeaker	930	–
Total	11,344	100

(14) *You know people dey speak different things.* **I con dey ah** but **na English everybody suppose dey speak.**
You know, people speak different languages. I figured, "ah, everyone's supposed to be speaking English."

(15) *Di baby no reach time* wey **I take born am.**
The baby came earlier than expected.

Projection of relation senses: Finally, we examined if the relation sense annotation in the ST can be projected to the TT, or if a shift in meaning occurred in the translation. We calculated this for dict project using the multi-label kappa metric. The intra-annotator agreement on relation sense in Nigerian Pidgin and English is high ($\kappa = .99$). This suggests that relation senses can be projected well and that little change in meaning occurred.

5 Corpus exploration

Table 4 presents the relation type distributions of the Nigerian Pidgin annotations. There is a roughly equal ratio of explicit to implicit relations in DiscoNaija, with the two relation types combined making up 88% of all data in the corpus. The remaining 12% consists mostly of EntRel and NoRel instances. Note that the Interspeaker relations were excluded from the percentage distribution since Interspeaker relations are not present in other corpora and would thus distort the comparison in later subsections.

In order to be able to interpret whether these patterns characterize discourse structure in Nigerian Pidgin, we need to compare these distributions to distributions from other language resources that contain similar genres. In the remainder of this section, we first present coarse-grained corpus distributions of DiscoNaija with other PDTB-inspired corpora, followed by a more in-depth comparison with a comparable corpus of spoken English, DiscoSPICE. The DiscoSPICE corpus consists of texts from the SPICE-Ireland corpus (Kallen & Kirk, 2008), with texts from broadcast interviews and telephone conversations. These genres are similar to the genres included in DiscoNaija, namely free conversations, broadcast reports from radio programs, life stories, comments on current state of affairs.

Table 5 The proportion of explicit relations versus implicit relations (and 95% confidence intervals) in PDTB 3.0 and various PDTB-based corpora containing spoken data

Corpus	Genre	# Relations	Prop. explicit	Prop. implicit
PDTB 3.0 (English)	Written, newspaper	53,631	.45 (CI: .45–.45)	.41 (CI: .41–.41)
TED-CDB (Chinese)	Spoken, prepared	15,540	.36 (CI: .35–.37)	.45 (CI: .44–.46)
TED-MDB (6 languages)	Spoken, prepared	3649	.41 (CI: .39–.43)	.35 (CI: .33–.37)
DiscoSPICE (English)	Spoken, spontaneous	1408	.64 (CI: .61–.66)	.13 (CI: .11–.15)
LUNA (Italian)	Spoken, spontaneous	1606	.66 (CI: .64–.68)	.30 (CI: .28–.32)
DiscoNaija (Nigerian Pidgin)	Spoken, spontaneous	11,344	.44 (CI: .43–.45)	.44 (CI: .43–.45)

5.1 Explicit and implicit relation types in DiscoNaija and other PDTB-inspired corpora

Table 5 presents a comparison between the PDTB 3.0, various PDTB-style corpora containing spoken data, and DiscoNaija in terms of the distribution of explicit vs. implicit relations. The rows do not add up to 100% because the table does not include the number of EntRel, NoRel and Hypophora relations. 95% confidence intervals for proportions were calculated using the Wilson score interval (with continuity correction).

The table shows that the corpora differ from each other in the proportion of relations that are marked explicitly. This could be attributed to various factors: there might be language-specific factors affecting the degree of marking, but the differences might also be due to the genres included in the corpora (e.g., prepared speeches in TED-CDB and TED-MDB, compared to spontaneous spoken language in DiscoSPICE and DiscoNaija).

Further, intra-sentential implicit relations were annotated in the PDTB 3.0 and TEDCDB, but not in DiscoNaija nor in the other corpora. Thus, a higher proportion of implicit relations should be expected in the PDTB and TED-CDB.

Despite not annotating intra-sentential implicit relations, DiscoNaija has a higher proportion of implicit relations when looking at the other corpora in the same genre of spontaneous speech. The proportion of implicit relations in Table 5 are calculated based on corpus size including NoRel, EntRel and Hypophora relations. A cleaner comparison of explicit-to-implicit ratio would be to calculate this excluding the other relation types. When doing so, DiscoNaija still has a higher proportion of implicit relations (DiscoNaija: 50%, CI: .49–.51; DiscoSPICE: 18%, CI: .15–.20; LUNA: 32%, CI: .29–.34). This might be attributed to language-specific differences. For example, the connective lexicon for Nigerian Pidgin is smaller compared to English and Italian. These distributions confirm our expectation that Nigerian Pidgin is characterized by a higher degree of implicit relations compared to other languages.

Table 6 Distribution of relations by level 2 relation senses (only those senses occurring > 5%)

Relation sense	# Total (Prop. of Corpus)	# Implicit (Prop. of Sense)
Contingency.Cause	2381 (.24)	1265 (.53)
Expansion.Conjunction	1935 (.19)	1013 (.52)
Temporal.Asynchronous	1165 (.12)	420 (.36)
Expansion.Level-of-detail	1039 (.10)	929 (.89)
Comparison.Concession	800 (.08)	282 (.35)
Contingency.Condition	691 (.07)	32 (.05)

Prop. of corpus represents the proportion of all instances in the corpus that received that relation sense label, and *prop. of sense* represents the proportion of instances that were implicit for that specific relation sense

5.2 Corpus distributions: DiscoNaija compared to DiscoSPICE

5.2.1 Relation senses

Table 6 shows the most frequently occurring level 2 relation senses, and their proportion of implicit relations. We see that Cause and Conjunction relations make up a large part of the corpus (together 43% of the corpus). The greatest disparity in the proportion of implicit relations is seen in Level-of-detail and Condition relations. As expected, the former are more likely to be implicit, whereas the latter are very likely to be marked. These trends replicate those found in written data in English (Asr & Demberg, 2012).

Since Nigerian Pidgin has fewer morphological markers for tense, we expected a lower rate of implicitness for Temporal Asynchronous relations. The PDTB level 2 sense Asynchronous includes both chronological and anti-chronological relations. In the absence of morphological markers, marking the temporal order is particularly relevant for antichronological relations, since these deviate from the real-world order of events (Münte et al., 1998; Scholman et al., 2022a; Ye et al., 2012). We therefore look at the implicitness rates for the level 3 types of Asynchronous relations in DiscoNaija and DiscoSPICE. Since DiscoSPICE contains relatively few instances of Asynchronous relations (87 instances in total), we also include PDTB 3.0 in the comparison. Figure 1 shows the results. Precedence relations tend to occur more than Succession relations in DiscoNaija compared to the other two corpora. However, there is one crucial difference: chronological precedence relations have a higher implicitness rate in DiscoNaija than in DiscoSPICE, but the anti-chronological succession relations tend to be expressed explicitly in DiscoNaija—that is, the succession implicitness proportion is 7% (CI: .04–.12) in DiscoNaija, 9% in DiscoSPICE (CI: .09–.37), and 15% (CI: .13–.17) in PDTB 3.0 (i.e., these proportions consider only the occurrences of succession relations). Taken together, these results support our expectation that anti-chronological relations are less likely to be expressed and more likely to be marked explicitly in Nigerian Pidgin.

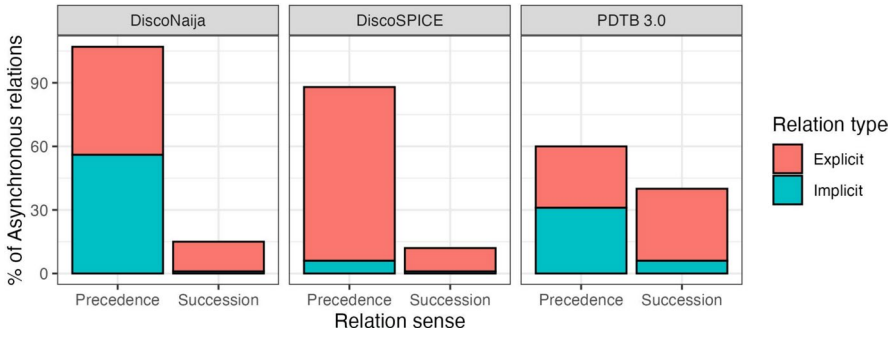

Fig. 1 Percentage of implicit and explicit Asynchronous subtypes in DiscoNaija and PDTB 3.0

When an anti-chronological succession relation occurred implicitly (n = 12), the temporal order of the events referred to in the arguments could typically be inferred using world knowledge, as is illustrated in (16): the pregnancy occurred before giving birth.

(16) *I just born now eh like seven months. Kai when I dey pregnant, e no dey easy o.*
I just gave birth uh… like seven months ago. Kai, when I was pregnant, it was not easy at all.
[Asynchronous.succession]

5.2.2 Connectives

5.2.2.1 NaijaLex 2.0 As part of the work described in the current study, we present an updated version of the Nigerian Pidgin connective lexicon NaijaLex. This update consists of various changes: during relation annotation, additional connective types were discovered; the possible relation senses that the connectives can express were updated from PDTB 2 senses to PDTB 3 senses; and the frequency of the connectives was updated. In this subsection, we therefore present new descriptive statistics for NaijaLex 2.0, which is made available online.[9] Table 7 presents descriptive statistics of NaijaLex 2.0. Note that variants are most commonly different spellings of a connective type (e.g., *cos* is a variant of the type *because*).

For each connective entry, the lexicon contains information on its frequency, alternative forms, syntactic category(/ies), English translation equivalents and non-connective usage. In addition, the various relation senses that the connective can signal are included, together with an example of the Nigerian Pidgin connective in every sense and the relation sense distribution.

No new Nigerian Pidgin connectives that do not originate from English were identified compared to NaijaLex 1.0. Thus, the lexicon still contains 18 connective

[9] https://osf.io/xns9z.

Table 7 Connectives in NaijaLex 2.0

Connective tokens	4952
Connective types	78
Connective variants	147
Connective types unique to Pidgin	18

types that are unique to Nigerian Pidgin, of which 9 are multi-word expressions. Examples of these are *upon sey* (English: *as*), *based on sey* (English: *since*) and *wey be sey* (English: *when/since/so that*). The fact that half of the 18 connective types unique to Nigerian Pidgin are multi-word expressions is particularly striking when contrasted with English, where many discourse connectives tend to be single lexical items (e.g., *because*, *although*, *however*, but note that English also has multi-word connectives such as *on the other hand*). This suggests that Nigerian Pidgin may construct discourse relations using phrases rather than compact, morphologically opaque words. This aligns with the general analytic nature of the language, where grammatical functions are often expressed through combinations of separate words rather than inflections or compact markers. Moreover, the presence of these multi-word connectives may also be related to the relatively young nature of Nigerian Pidgin as a language. Connective functions are often realized first through periphrastic means (i.e., using several words), and only later become lexicalized into single-word items through grammaticalization (Zufferey & Degand, 2024). English connectives like *although* or *because* have gone through long historical processes of fusion and reduction from earlier multi-word expressions (e.g., *all though*, *by cause*), whereas Nigerian Pidgin may still be in earlier stages of that trajectory.

5.2.2.2 Discourse marking in Nigerian Pidgin While DiscoNaija contains fewer unique connective types than PDTB 3.0 (n = 173), it does contain more unique connective types compared to DiscoSPICE (n = 48). The difference between these three corpora lies in their genre and size: certain connectives (e.g., *notwithstanding*) are more likely to occur in written, formal text such as the text in PDTB 3.0 than in spoken spontaneous text, and DiscoSPICE is smaller than DiscoNaija, which might lead to less frequent connectives not occurring in DiscoSPICE.

Given that Nigerian Pidgin is characterized by serial verb constructions, we expected that coordinating conjunctions would be less frequent in Nigerian Pidgin compared to English. Table 8 shows that NaijaLex contains occurrences of 6 of the 7 main connectives in the syntactic category coordinating conjunction (additionally, NaijaLex includes a Nigerian Pidgin coordinating conjunction, *abi*). It appears that these coordinating connectives indeed occur less frequently in DiscoNaija than in PDTB 3.0: 37% (CI: .36–.39) of all connective annotations consist of coordinating connectives, versus 61% (CI: .60–.61) in PDTB 3.0. These results are in line with our expectation that coordinating connectives are less frequent in Nigerian Pidgin than in English. The difference in connective frequency between DiscoNaija and PDTB is particularly big for *and*; (17) presents an example of such a case where *and* is added in translation.

Table 8 Occurrences of coordinating connectives (and percentage of all connectives) in DiscoNaija and PDTB 3.0

Connective	DiscoNaija count (Prop.)	PDTB 3.0 count (Prop.)
and	539 (13%)	8252 (34%)
but	433 (10%)	4498 (19%)
for	7 (0%)	69 (0%)
nor	0 (0%)	33 (0%)
or	45 (1%)	397 (2%)
so	542 (13%)	1304 (5%)
yet	1 (0%)	152 (1%)
Total	1567 (37%)	14,705 (61%)

Table 9 Ten most frequent connectives in DiscoNaija and their percentage out of all explicit relations

Connective	Count	%
if	545	13
so	539	13
and	533	13
but	425	10
because	355	8
con (from English 'come')	299	7
when	256	6
as	246	6
den (English: 'then')	125	3
before	80	2

(17) *Na so all of us pack oursef* **go village.**
So all of us packed our things and went to the village.
[Conjunction]

Table 9 presents the top 10 most frequent connectives and their proportion of occurrence in the explicit data. The ranking is different from DiscoSPICE, where the three most frequent explicit connectives are *and*, *but* and *so*. The difference in occurrence of *but* might be due to Comparison relations being relatively infrequent in DiscoNaija. The difference in occurrence of *if* might be due to the fact that the DiscoNaija texts contain descriptions of recipes, in which conditional relations were very common (see (18)).

(18) If **di meat never soft,** *you put am dat period wey be sey you dey put all your ingredients* so dat everyting go boil together.
If the meat is not soft yet, you add it when you put all your ingredients so that everything boils together.
[Condition]

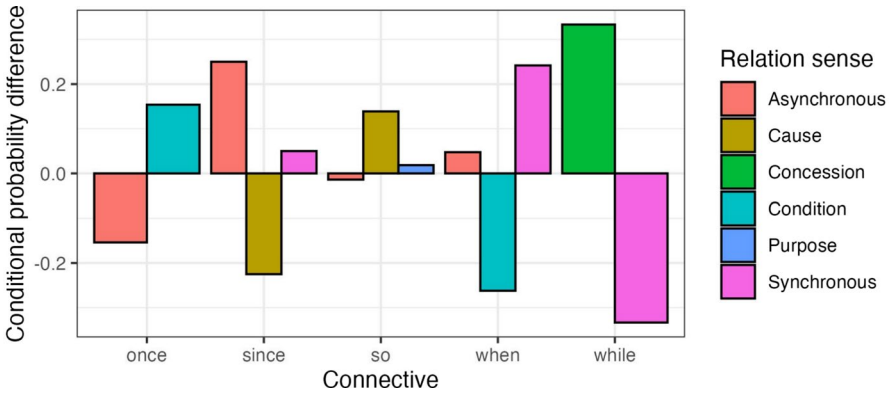

Fig. 2 Conditional probability differences of discourse relations given the top five most divergent connectives between Nigerian Pidgin and English (DiscoSPICE). Positive values indicate a higher probability of a connective occurring with a given relation in Nigerian Pidgin compared to English, and negative values indicate a lower probability

Despite shared origins, some connectives in Pidgin might have broadened or shifted meaning compared to their English origins. Our dataset allows us to investigate whether the discourse signals associated with connectives originally stemming from English have diverged or remained consistent across the two languages. Specifically, we compare the conditional probabilities (distribution of discourse senses per connective, P(sense|connective)) between connectives in DiscoNaija and DiscoSPICE to study linguistic change.

Figure 2 presents the difference in conditional probabilities for the five connectives of English origin with the largest divergence. First, the connective *once* tends to be used more in a condition sense in Pidgin, whereas in English, it tends to be used more in a temporal asynchronous sense. This suggests that *once* in Pidgin has extended beyond its English temporal meaning to take on a more conditional function, possibly compensating for the more limited range of dedicated conditional signals in the language (e.g., tense as a conditional signal). By contrast, the connective *since* is more strongly associated with temporal asynchronous relations in Pidgin, but with cause relations in English than in Nigerian Pidgin. This suggests that the causal function of *since* has weakened in Pidgin, potentially also to make up for the lack of morphological markers to express temporal ordering in Nigerian Pidgin. The connective *so* is more strongly associated with cause in Pidgin than in English. *When* is used more variably in Nigerian Pidgin (not only to express condition, but also temporal synchronous and asynchronous), possibly reflecting a broader semantic range for the connective in the language. Finally, *while* tends to signal Synchronous relations in English, but this function is less prominent in Nigerian Pidgin. The lower frequency of *while* for synchronous marking might reflect a preference for paratactic structures or alternative syntactic strategies to express simultaneity in Pidgin discourse. These findings indicate that while some connectives retain their discourse functions across

Table 10 Five most frequently occurring connectives that are unique to Nigerian Pidgin in DiscoNaija

Connective	Count	Percentage (%)
con	302	7.2
naim	79	1.9
make	39	0.9
like sey	26	0.6
sey	26	0.6

both languages, others exhibit shifts, likely due to differences in grammatical structures, the multifunctionality of discourse markers in Pidgin, and the need to compensate for gaps in the lexicon.

5.2.2.3 Connectives unique to Nigerian Pidgin In total, 13% of all connective occurrences in DiscoNaija are Nigerian Pidgin connectives not existing in English. Table 10 presents the five most frequent connectives that are not English cognates. We will elaborate on the meanings and origins of each of these connectives.

Con has evolved from the English verb 'come' and can now be used as an auxiliary, in which case it is frequently translated as connectives expressing temporal and cause relations in English, as in (19) and (20). It is used frequently in narration to connect the events in one utterance with the following events in the next utterance.

(19) *I say ah! I con realise sey omo na dis Pidgin na im make us connect like dat.*
I said "ah!". Then I realized that, wow, this Pidgin brought us together.
[Precedence]

(20) *Toh I no get money,* **I** con **s-… sell dem five, five hundred.**
Well, I don't have money, so I s-… sold them for five hundred naira each.
[Result]

The connective *naim* is a contraction of the auxiliary verb *na*, which is often used as a focus particle, and the 3SG pronoun *im*. It has grammaticalized and now functions as a cause or temporal connective, equivalent to 'so' or 'then', as in (21) and (22).

(21) *I say bring am now!* Naim **de wrap am bring am.**
I said, "bring it here." So they wrapped it up and brought it to me.
[Result]

(22) *Im say e wan tell us one story, make all of us listen.* Naim **e con start.**
He said he wanted to tell us a story, and all of us should listen. Then he started.
[Precedence]

Similar to *con*, *make* originated from an English verb, and can function as a connective when used as an auxiliary. In a non-connective usage, it often functions as a directive, its meaning equivalent to 'should' in English. This causal

function has likely extended to its usage as a causal discourse connective. It is frequently translated as 'so' or 'so that' in English, as in (23).

(23) She no go resign ni, *she no go go house ni,* make **she go rest now.**
Won't she resign? Won't she go home so she can go and rest?
[Arg2-as-goal]

In Nigerian Pidgin, the word *sey* evolved from English 'say' but took on the grammatical function of a complementizer rather than a verb. Its usage is similar to the English word 'that' in reported speech or indirect statements. In DiscoNaija, *sey* is translated as a causal or temporal connective, similar to English 'because', 'since', 'if' and 'when', as in (24). The connective *like sey* is a combination of the English connective 'like' and the Pidgin subordinating conjunction *sey*, and is commonly used to express similarity and manner relations, as in (25).

(24) Or e go tell you sey *meh you waka go find your own,* sey **im ma, don find im own.**
Or he will ask you to go find your own blessings, since he found his own.
[Reason]

(25) In fact, she was looking, she dey look, *she look so miserable* like sey **ehe dat wish na punishment**, because she no get choice.
In fact, she was looking… she looked… she looked so miserable like it was a punishment because she didn't have a choice.
[Similarity, Arg2-as-manner]

6 Discussion

We presented DiscoNaija, a freely available corpus annotated with PDTB-style discourse relations. DiscoNaija consists of an annotation layer on the Naija Treebank, which is a corpus of transcribed Pidgin texts and translated English texts (Caron et al., 2019). The genre of the texts included can be classified as spontaneous spoken discourse, including dialogues and monologues on a variety of topics and uttered by a variety of speakers.

The corpus was created by first annotating the English translated text, and then projecting these annotations to the Nigerian Pidgin text. The DiscoNaija corpus thus contains discourse annotations for both languages. We assessed the applicability of the proposed annotation projection method, which consisted of first aligning the arguments and the connectives (if present), and then projecting and updating the relation type and relation sense if need be. The paper provided agreement statistics that demonstrate the reliability of the annotations, both within a language and between languages.

Based on the syntactic distributions of Nigerian Pidgin, as well as characteristics typical of spoken discourse, we made a few adaptations to the PDTB annotation approach. We added the relation type Interspeaker, for those adjacent sentences that were not uttered by the same person. Further, we added a feature to the dataset

to mark disfluent arguments (i.e., when one or both utterances of a relation was interrupted).

We formulated several expectations regarding how discourse structure is realized in Nigerian Pidgin and how this might compare to English. As expected, the corpus distributions showed that DiscoNaija contains a higher proportion of implicit discourse relations compared to corpora of spontaneous spoken discourse in English and Italian. Moreover, the corpus study also revealed that anti-chronological temporal relations are more likely to be expressed explicitly in Nigerian Pidgin compared to English, which we expect is due to Nigerian Pidgin having fewer morphological tense markers. Further, we found that coordinating conjunctions occur less in Nigerian Pidgin than in English.

The development of the Nigerian Pidgin discourse-annotated corpus described in the current work addresses the gap in the field in terms of pidgin language resources focusing on discourse structural features. Such datasets are crucial for training classifiers that can automatically uncover discourse relations in a text (a task referred to as discourse parsing), which in turn supports down-stream tasks such as argument mining (Kirschner et al., 2015), summarization (Dong et al., 2021; Xu et al., 2020) and relation extraction (Tang et al., 2021). Discourse relation classifiers need large amounts of training data to perform accurately. Translating English datasets into Nigerian Pidgin would not necessarily suffice or erase the need for original resources, since genre and cultural concepts play a role in NLP (discourse) tools as well (Lent et al., 2024; Scholman et al., 2021). DiscoNaija can therefore be a valuable source for future research efforts in training NLP tools.

The corpus offers possibilities for various research purposes. First, it can be used to promote further development on discourse relation recognition and discourse-level NLP tasks. In fact, Saeed et al. (2025) leveraged the implicit relation annotations in DiscoNaija as a test set for evaluating several automatic relation classification methods. Their study explored several alternative setups: (1) applying an English discourse classifier directly to Nigerian Pidgin; (2) translating the Pidgin text into English, classifying the relations, and projecting the results back onto the original Nigerian Pidgin text; (3) training a dedicated Nigerian Pidgin classifier on synthetic discourse relation annotations using a data-augmentation approach. This model achieved accuracy/F1 scores of 0.631/0.461 for 4-way relation sense classification and 0.440/0.327 for 11-way classification, outperforming the other two approaches.

These findings underscore the need for more discourse-annotated data in low-resource languages: it is highly likely that even better results could be achieved if additional data was available for training such a parser on high-quality labelled data. While DiscoNaija contributes to this goal (with over 11,000 annotations), it remains insufficient for training robust neural discourse parsers, especially when a portion must be held out for reliable evaluation. As the field progresses, continued annotation efforts and training strategies will be key to advancing discourse-aware NLP for Nigerian Pidgin and related languages.

Second, the corpus contains parallel texts: manual translations of the Nigerian Pidgin transcribed text into English. DiscoNaija contains annotations on both the Nigerian Pidgin and the English texts, and the dataset can thus be used for studying

translation effects on discourse relation interpretations. The English part of the dataset can also be used as additional training data for English discourse NLP tools, representing a spontaneous spoken genre (a domain for which not many discourse-annotated resources are available in English).

Third, the corpus allows us to study features that characterize Nigerian Pidgin discourse structure, and thus how discourse coherence can be expressed in creole languages. A first effort to do so was made in the current paper. These findings raise questions regarding the cognitive processing of the Nigerian Pidgin language and by Nigerian Pidgin (monolingual and bilingual) speakers. For example, given that relations are more often marked implicitly in Nigerian Pidgin, a possible hypothesis is that comprehenders might rely less on connectives during interpretation compared to comprehenders of other languages, and thus there might be less of a facilitative effect of the connective. Future work can also focus on possible differences in connective usage between monolingual Pidgin and bilingual Pidgin-English speakers: do bilingual speakers tend to produce a greater proportion of explicit discourse relations in Pidgin speech (i.e., a possible transfer effect from their English language statistics)?

In sum, the main contributions of this paper have been (i) the presentation of DiscoNaija—a parallel discourse-annoted corpus of Nigerian Pidgin and English spoken spontaneous conversations and monologues, (ii) the update of NaijaLex 2.0—an existing connective lexicon of Nigerian Pidgin, (iii) the evaluation of an annotation projection approach, and (iv) an initial analysis of discourse relations and connective distributions that are characteristic of Nigerian Pidgin. We hope the resources presented here will be used to spur future research on Nigerian Pidgin in the computational and (psycho-)linguistic fields.

Acknowledgements We are highly grateful to Emeka Felix Onwuegbuzia for his valuable insights on Nigerian Pidgin.

Author contributions MS and VD were responsible for conceptualisation, funding acquisition and supervision. MS, MM and AB were responsible for data curation. MS and MM were responsible for formal analysis and wrote the original draft of the manuscript. All authors contributed to the editing of the manuscript.

Funding This research was funded by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074).

Data availability The corpus is available at https://osf.io/8m5vk/.

Declarations

Competing interests The authors declare no competing interests.

References

Ahmad, W. U., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., & Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. *Proceedings of NAACL-HLT* (pp. 2440–2452).

Aikhenvald, A. Y., & Dixon, R.M. (2005). *Serial verb constructions: A cross-linguistic typology*. OUP Oxford.

Al-Saif, A., & Markert, K. (2010). The Leeds Arabic discourse treebank: annotating discourse connectives for Arabic. *Proceedings of the 7th international conference on language resources and evaluation (LREC)* (pp. 2046–2053). Valletta, Malta.

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. *Proceedings of the international conference on computational linguistics (COLING)* (pp. 2669–2684). Mumbai, India.

Bakker, P. (2008). Pidgins versus creoles and pidgincreoles. In S. Kouwenberg & J. V. Singler (Eds.), *The handbook of pidgin and creole studies* (pp. 130–157). Wiley-Blackwell.

Becher, V. (2011). When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies, 23*(1), 26–47. https://doi.org/10.1075/target.23.1.02bec

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences, 26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T. J., & Stede, M. (2018). Constructing a lexicon of Dutch discourse connectives. *Computational Linguistics in the Netherlands Journal, 8*, 163–175.

Bourgonje, P., & Lin, P.-J. (2024). Projecting annotations for discourse relations: Connective identification for low-resource languages. *Proceedings of the 5th workshop on computational approaches to discourse (CODI 2024)* (pp. 39–49).

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press.

Carlson, L., & Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, *54*, 1–56,

Caron, B., Courtin, M., Gerdes, K., & Kahane, S. (2019). A surface-syntactic UD treebank for Naija. *TLT 2019, Treebanks and Linguistic Theories, Syntaxfest.*

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy* (pp. 35–54).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Courtin, M., Caron, B., Gerdes, K., Kahane, S. (2018). Establishing a language by annotating a corpus. *anndh 2018 annotation in digital humanities* (Vol. 2155, pp. 7–11).

Crible, L., Abuczki, A., Burksaitiene, N., Furkó, P., Nedoluzhko, A., Rackeviciene, S., Oleškevičienė, G. V., & Zikánová, S. (2019). Functions and translations of discourse markers in TED talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics, 142*, 139–155. https://doi.org/10.1016/j.pragma.2019.01.012

Crible, L., & Cuenca, M.-J. (2017). Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse, 8*(2), 149–166. https://doi.org/10.5087/dad.2017.207

Crible, L., & Zufferey, S. (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. *Proceedings of the 11th Joint ACL-ISO workshop on interoperable semantic annotation (ISA-11).*

Das, D., Stede, M., Ghosh, S. S., & Chatterjee, L. (2020). DiMLex-Bangla: A lexicon of Bangla discourse connectives. *Proceedings of the 12th language resources and evaluation conference* (pp. 1097–1102).

DeGraff, M. (1999). *Language creation and language change: Creolization, diachrony, and development*. MIT Press (MA).

Dong, Y., Mircea, A., & Cheung, J. C. K. (2021). Discourse-aware unsupervised summarization for long scientific documents. *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume* (pp. 1089–1102).

Dou, Z.-Y., & Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. *Conference of the European chapter of the Association for Computational Linguistics (EACL).*

Esizimetor, D. O. (2009). What orthography for naijá? *Proceedings of the conference on Nigerian Pidgin, University of Ibadan, Nigeria.*

Faraclas, N. (2004). Nigerian Pidgin English: Morphology and syntax. In B. Kortmann & E. W. Schneider (Eds.), *A handbook of varieties of English* (pp. 2020–2045). De Gruyter Mouton. https://doi.org/10.1515/9783110197181-121

Faraclas, N. (2013). Nigerian pidgin [Bibliographical record]. S. M. Michaelis, P. Maurer, M. Haspelmath, & M. Huber (Eds.), *The survey of pidgin and creole languages. Vol. I: English-based and Dutch-based languages* (pp. 176–184). Oxford University Press.

Faraclas, N. (2021). Naija: A language of the future. *Current Trends in Nigerian Pidgin English: A Sociolinguistic Perspective, 117*, 9. https://doi.org/10.1515/9781501513541

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science, 3*(1), 67–90. https://doi.org/10.1207/s15516709cog03014

Hoek, J., Evers-Vermeul, J., & Sanders, T. J. M. (2015). The role of expectedness in the implicitation and explicitation of discourse relations. *Proceedings of the second workshop on discourse in machine translation (DiscoMT)* (pp. 41–46).

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. M. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics, 121*, 113–131. https://doi.org/10.1016/j.pragma.2017.10.010

Igboanusi, H. (2008). Empowering Nigerian Pidgin: A challenge for status planning? *World Englishes, 27*(1), 68–82. https://doi.org/10.1111/j.1467-971X.2008.00536.x

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 6282–6293).

Kallen, J. L., & Kirk, J. M. (2008). *ICE-Ireland: A user's guide: Documentation to accompany the Ireland Component of the International Corpus of English (ICE-Ireland)*. Cl´o Ollscoil na Banr´ıona.

Kirschner, C., Eckle-Kohler, J., & Gurevych, I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. *Proceedings of the 2nd workshop on argumentation mining* (pp. 1–11).

Kishimoto, Y., Sawada, S., Murawaki, Y., Kawahara, D., & Kurohashi, S. (2018). Improving crowdsourcingbased annotation of Japanese discourse relations. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).*

Kurfalı, M., Ozer, S., Zeyrek, D., & Mendes, A. (2020). Ted-MDB lexicons: TrEnConnLex, PtEnConnLex. *Proceedings of the first workshop on computational approaches to discourse* (pp. 148–153).

Laali, M., & Kosseim, L. (2014). Inducing discourse connectives from parallel texts. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 610–619).

Laali, M., & Kosseim, L. (2017). Improving discourse relation projection to build discourse annotated corpora. *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 407–416).

Lapshinova-Koltunski, E., Pollkläsener, C., & Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics, 119*, 5–22. https://doi.org/10.14712/00326585.020

Lent, H., Bugliarello, E., De Lhoneux, M., Qiu, C., & Søgaard, A. (2021). On language models for Creoles. *Proceedings of the 25th conference on computational natural language learning* (pp. 58–71).

Lent, H., Ogueji, K., de Lhoneux, M., Ahia, O., & Søgaard, A. (2022). What a creole wants, what a creole needs. *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6439–6449).

Lent, H., Tatariya, K., Dabre, R., Chen, Y., Fekete, M., Ploeger, E., Zhou, L., Armstrong, R.-A., Eijansantos, A., Malau, C., Heje, H. E., Lavrinovics, E., Kanojia, D., Belony, P., Bollmann, M., Grobol, L., de Lhoneux, M., Hershcovich, D., DeGraff, M., … Bjerva, J. (2024). CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics, 12*, 950–978. https://doi.org/10.1162/tacl_a_00682

Lin, P.-J., Saeed, M., Chang, E., & Scholman, M. C. (2023). Low-resource cross-lingual adaptive training for Nigerian Pidgin. *Interspeech 2023* (pp. 3954–3958).

Lin, P.-J., Scholman, M. C., Saeed, M., & Demberg, V. (2024). Modeling orthographic variation improves NLP performance for Nigerian pidgin. *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 11510–11522).

Long, W., Webber, B., & Xiong, D. (2020). TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 2793–2803).

Marchal, M., Scholman, M. C., & Demberg, V. (2021). Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin. *Proceedings of the 2nd workshop on computational approaches to discourse* (pp. 84–94).

Marchal, M., Scholman, M. C., Yung, F., & Demberg, V. (2022, October). Establishing annotation quality in multi-label annotations. *Proceedings of the 29th international conference on computational linguistics* (pp. 3659–3668). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2022.coling-1.322

Mensah, E., Ukaegbu, E., & Nyong, B. (2021). Towards a working orthography of Nigerian Pidgin. *Current Trends in Nigerian Pidgin English*. https://doi.org/10.1515/9781501513541

Meyer, T., Popescu-Belis, A., Zufferey, S., & Cartoni, B. (2011). Multilingual annotation and disambiguation of discourse connectives for machine translation. *Proceedings of the SIGDIAL 2011 conference* (pp. 194–203).

Mírovsky, J., Synková, P., & Poláková, L. (2021). Extending coverage of a lexicon of discourse connectives using annotation projection. *The Prague Bulletin of Mathematical Linguistics, 117*, 5–26. https://doi.org/10.14712/00326585.015

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature, 395*(6697), 71–73. https://doi.org/10.1038/25731

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. *Proceedings of the 1st workshop on multilingual representation learning* (pp. 116–126).

Ojarikre, A. (2013). Perspectives and problems of codifying Nigerian Pidgin English orthography. *Perspectives, 3*(12), 126–133.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., & Joshi, A. (2009). The Hindi Discourse Relation Bank. *Proceedings of the third linguistic annotation workshop (LAW III)* (pp. 158–161).

Parkvall, M. (2008). The simplicity of creoles in a cross-linguistic perspective. *Language complexity* (pp. 265–285). John Benjamins.

Ponti, E. M., Glavas, G., Majewska, O., Liu, Q., Vulic, I., & Korhonen, A. (2020). Xcopa: A multilingual dataset for causal commonsense reasoning. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 2362–2376).

Prasad, R., McRoy, S., Frid, N., Joshi, A., & Yu, H. (2011). The Biomedical Discourse Relation Bank. *BMC Bioinformatics, 12*(1), 1–18. https://doi.org/10.1186/1471-2105-12-188

Prasad, R., Webber, B., & Joshi, A. (2014). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics, 40*(4), 921–950. https://doi.org/10.1162/COLI_a_00204

Prasertsom, P., Jaroonpol, A., & Rutherford, A. T. (2024). The Thai discourse treebank: Annotating and classifying Thai discourse connectives. *Transactions of the Association for Computational Linguistics, 12*, 613–629. https://doi.org/10.1162/tacl_a_00650

Rehbein, I., Scholman, M. C., & Demberg, V. (2016, May). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1039–1046). Portoroˇz, Slovenia: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L161165

Rohde, H., Dickinson, A., Schneider, N., Clark, C., Louis, A., & Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. *Proceedings of the 10th linguistic annotation workshop (LAW X)* (pp. 49–58). Berlin, Germany.

Ruder, S., Vulic, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research, 65*, 569–631. https://doi.org/10.1613/jair.1.11640

Saeed, M. Y. G. S., Bourgonje, P., & Demberg, V. (2025). Implicit discourse relation classification for Nigerian pidgin. *Proceedings of the 31st international conference on computational linguistics* (pp. 2561– 2574).

Sanders, T. J. M., Spooren, W. P. M. S., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*(1), 1–35. https://doi.org/10.1080/01638539209544800

Schiffrin, D. (1987). *Discourse markers* (No. 5). Cambridge University Press.

Scholman, M. C., Blything, L., Cain, K., Hoek, J., & Evers-Vermeul, J. (2022a). Discourse rules: The effects of clause order principles on the reading process. *Language, Cognition and Neuroscience, 37*(10), 1277–1291. https://doi.org/10.1080/23273798.2022.2077971

Scholman, M. C., Dong, T., Yung, F., & Demberg, V. (2021). Comparison of methods for explicit discourse connective identification across various domains. *Proceedings of the 2nd workshop on computational approaches to discourse* (pp. 95–106).

Scholman, M. C., Dong, T., Yung, F., & Demberg, V. (2022b, June). DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. *Proceedings of the thirteenth international conference on language resources and evaluation (LREC'22).* Marseille, France: European Language Resources Association (ELRA).

Scholman, M. C., Pyatkin, V., Yung, F., Dagan, I., Tsarfaty, R., & Demberg, V. (2022c). Design choices in crowdsourcing discourse relation annotations: The effect of worker selection and training. *Proceedings of the thirteenth language resources and evaluation conference* (pp. 2148–2156).

Siegel, J. (2008). *The emergence of pidgin and creole languages*. Oxford University Press.

Sluyter-Gäthje, H., Bourgonje, P., & Stede, M. (2020). Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. *Proceedings of the twelfth language resources and evaluation conference* (pp. 1044–1050).

Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory, 6*, 241–266. https://doi.org/10.1515/cllt.2010.009

Tang, J., Lin, H., Liao, M., Lu, Y., Han, X., Sun, L., Xie, W., & Xu, J. (2021). From discourse to narrative: Knowledge projection for event relation extraction. *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long Papers)* (pp. 732–742).

Thomason, S. G. (2001). *Language contact: An introduction*. Edinburgh University Press.

Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. K. (2010). Annotation of discourse relations for conversational spoken dialogs. *LREC*.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. *Proceedings of workshop on annotation and exploitation of parallel corpora (AEPC)* (pp. 83–82).

Wang, J., & Lan, M. (2015, July). A refined end-to-end discourse parser. *Proceedings of the nineteenth conference on computational natural language learning - shared task* (pp. 17–24). Association for Computational Linguistics. Retrieved from https://aclanthology.org/K15-2002

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). *The Penn Discourse Treebank 3.0 annotation manual*. University of Pennsylvania.

Xu, J., Gan, Z., Cheng, Y., & Liu, J. (2020). Discourse-aware neural extractive text summarization. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5021–5031).

Ye, Z., Kutas, M., George, M. S., Sereno, M. I., Ling, F., & Münte, T. F. (2012). Rearranging the world: Neural network supporting the processing of temporal connectives. *NeuroImage, 59*(4), 3662–3667. https://doi.org/10.1016/j.neuroimage.2011.11.039

Yung, F., Scholman, M. C., Lapshinova-Koltunski, E., Pollkläsener, C., Demberg, V. (2023, September). Investigating explicitation of discourse connectives in translation using automatic annotations. In S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, & M. Alikhani (Eds.), *Proceedings of the 24th annual meeting of the special interest group on discourse and dialogue* (pp. 21–30). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.sigdial-1.2

Yung, F., Scholman, M. C., Zikánová, S., & Demberg, V. (2024). DiscoGeM 2.0: A parallel corpus of English, German, French and Czech implicit discourse relations. *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 4940–4956).

Zeyrek, D., & Er, M. E. (2022). A description of Turkish Discourse Bank 1.2 and an examination of common dependencies in Turkish Discourse. arXiv preprint arXiv:2207.05008. https://doi.org/10.48550/arXiv.2207.05008

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation, 54*(2), 587–613. https://doi.org/10.1007/s10579-019-09445-9

Zeyrek, D., Mendes, A., Oleskeviciene, G. V., & Özer, S. (2022). An exploratory analysis of TED talks¨ in English and Lithuanian, Portuguese and Turkish Translations: Results from the analysis of an annotated multilingual corpus. *Contrastive Pragmatics, 3*(3), 452–479. https://doi.org/10.1163/26660393-bja10052

Zhou, Y., & Xue, N. (2012). PDTB-style discourse annotation of Chinese text. *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 69–77).

Zikánová, S., Mírovsky, J., & Synková, P. (2019). Explicit and implicit discourse relations in the Prague discourse treebank. *Text, speech, and dialogue: 22nd International conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22* (pp. 236–248).

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies, 26*(3), 361–384. https://doi.org/10.1075/target.26.3.02zuf

Zufferey, S., & Degand, L. (2024). *Connectives and discourse relations*. Cambridge University Press.