Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin

Marian Marchal[1], Merel Scholman[1], and Vera Demberg[1,2]
[1]Language Science and Technology, [2]Computer Science, Saarland University
Saarbrücken, Germany
{marchal,m.c.j.scholman,vera}@coli.uni-saarland.de

Abstract

Cross-linguistic research on discourse structure and coherence marking requires discourse-annotated corpora and connective lexicons in a large number of languages. However, the availability of such resources is limited, especially for languages for which linguistic resources are scarce in general, such as Nigerian Pidgin. In this study, we demonstrate how a semi-automatic approach can be used to source connectives and their relation senses and develop a discourse-annotated corpus in a low-resource language. Connectives and their relation senses were extracted from a parallel corpus combining automatic (PDTB end-to-end parser) and manual annotations. This resulted in Naija-Lex, a lexicon of discourse connectives in Nigerian Pidgin with English translations. The lexicon shows that the majority of Nigerian Pidgin connectives are borrowed from its English lexifier, but that there are also some connectives that are unique to Nigerian Pidgin.

1 Introduction

Resources such as discourse-annotated corpora (e.g. PDTB, Prasad et al., 2008) and connective lexicons (e.g. DimLEX, Stede and Umbach, 1998) have proven highly valuable in investigating discourse structure and coherence marking, but they are only available for a limited number of languages. One promising way to expand these resources is to project annotations from languages for which such resources are available to under-resourced languages. The current paper sets out to investigate how automatic discourse parsing can be combined with manual annotation to construct a connective lexicon in a low-resource language. We apply this method to create Naija-Lex, a lexicon of Nigerian Pidgin, but the method can also be adapted to apply to other low-resource languages.

Pidgin and creole languages, such as Nigerian Pidgin (commonly referred to as 'Pidgin' or 'Naija'), tend to be less complex on a syntactic and morphological level than their lexifiers (Szmrecsanyi et al., 2009). This raises interesting questions with respect to their discourse complexity, such as: how is discourse structure realized in contact languages? And how does this affect discourse marking? In addition, because pidgins are still in an evolutionary stage, they are a testing ground for examining how connectives emerge. However, the scarcity of linguistic resources, such as discourse-annotated corpora, for these languages makes it difficult to answer these questions. Another contribution of this paper is the creation of a discourse-annotated layer for a parallel corpus of Nigerian Pidgin, which can be used to study these questions.

We adopt a semi-automatic approach using automatic connective identification and annotation projection, combined with manual annotation, to initialize a list of discourse connective candidates in Nigerian Pidgin. This allows us to evaluate how existing tools can be adopted to investigate low-resource languages. More specifically, we assess the usefulness of the PDTB end-to-end parser (Lin et al., 2014) in identifying connectives in informal spoken English translated text, to source connectives in the target language.

The main contributions of the present research are the following:

- We identify challenges in constructing a lexicon in an under-resourced language and present a methodology for sourcing discourse connectives and their associated relation senses in such a language.
- We create a lexicon of discourse connectives in Nigerian Pidgin with English translations (Naija-Lex), which enables further research on discourse structure and coherence marking in Pidgin.
- We add a discourse-annotated layer to an existing Nigerian Pidgin-English parallel corpus

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 84–94
November 10–11, 2021.

containing explicit connectives and their relation senses.

Section 2 discusses earlier work on automatic discourse annotation and lexicon construction, after which Nigerian Pidgin is introduced in Section 3. Section 4 describes the method used to identify the connectives in Pidgin. We elaborate on the NaijaLex lexicon that we created using this method in Section 5 and evaluate our method in Section 6.

2 Related work

2.1 Lexicon creation

Connective inventories have been developed for various languages, including German (Stede and Umbach, 1998), French (Roze et al., 2012), Chinese (Zhou and Xue, 2015) and English (Das et al., 2018), among others (see also Stede et al., 2018). Recently, these efforts have been extended with several multi-lingual connective databases (Bourgonje et al., 2017; Kurfalı et al., 2020).

One challenging aspect about constructing a connective lexicon for a new language is defining the category of connectives. Traditionally, the PDTB (Prasad et al., 2008) has placed strong syntactic restrictions on which lexical items can function as connectives, considering only subordinating and coordinating conjunctions and adverbials. Stede et al. (2018) point out that for multilingual lexicons, such a strict definition is not feasible, as languages differ highly in the way coherence relations are expressed. Still, Stede et al. (2018) restrict the class of connectives syntactically, by posing that a connective cannot be inflected or modified, and that verbs should be excluded for this reason.

Similar to certain other lexicons (e.g. LDM-PT, Mendes and Lejeune, 2016), the lexicon we present here will maintain less strict requirements on the syntactic category and modifiability of the lexical item. In this way, we can allow for more variation that might be present in a relatively young language and obtain a more complete overview of discourse marking in Nigerian Pidgin.

2.2 Connective identification

In the last decade, shallow discourse parsing of explicit connectives has received considerable attention in the computational discourse community (e.g. Lin et al., 2014; Pitler and Nenkova, 2009; Wang and Lan, 2015; Zeldes et al., 2019). One challenge for such automatic parsers is that a discourse connective often also has a non-connective function, which is difficult to distinguish. Still, automatic discourse parsers for English achieve an F1 of almost 95% in identifying explicit connectives (Lin et al., 2014; Wang and Lan, 2015).

Discourse parsers are still very much restricted in terms of language and domain, because of a strong focus on the use of the PDTB for training. Parsers on languages other than English often reach much lower performance (e.g. Xue et al., 2016), although the recent DisRPT connective detection task (Zeldes et al., 2019) has yielded some high performing systems in other languages as well (Muller et al., 2019; Yu et al., 2019). However, these parsers still require training data, which is not available for low-resource languages.

Using parallel corpora might therefore be a promising method to identify connectives and then project their annotations to a low-resource language. One possible difficulty is that these corpora might be from different domains than what the parsers have been trained on, which leads to lower performance (Ramesh and Yu, 2010; Knaebel and Stede, 2020; Scholman et al., 2021).

We will use the PDTB end-to-end parser (Lin et al., 2014), as it has been shown to perform better on spontaneous spoken data than other parsers (Scholman et al., 2021). However, performance of the parser drops considerably when comparing spoken language with the newspaper text data it was trained on. We will therefore also assess its reliability on an out-of-domain translated text.

2.3 Connective projection

Several previous studies have used research available on connectives in one language (mostly English) to identify connectives in another language. Zhou and Xue (2012) and Das et al. (2020) have used bilingual dictionaries to translate English connectives to Chinese and Bangla respectively as part of identifying connectives in these target languages to construct a lexicon. However, no dictionaries are available for Nigerian Pidgin.

Another line of research in using cross-lingual information to source connectives is to use parallel corpora. Versley (2010) projected discourse connectives across an English-German parallel corpus, but found that simple word alignment is not enough to reliably identify connectives in text and that performance increased when bootstrapping using an established connective list. Bourgonje et al. (2018) combined a German connective lexicon with paral-

lel text corpora to find Dutch connectives and subsequently manually checked and annotated them for relation sense. However, this approach does not automatically distinguish between connective and non-connective usages of connectives in the source languages, which always needs to be resolved manually. The current research attempts to resolve this by using a shallow discourse parser to identify connectives in the translation text automatically.

A more similar approach to that adopted in the current study is that of Laali and Kosseim (2014). They used the PDTB end-to-end parser to identify English discourse connectives in the Europarl corpus source text, after which they used word alignment to identify connectives in the parallel French texts. This mapping was further improved with additional filtering based on frequency and syntactic information. However, performance is highly dependent on the length and syntactic categories of discourse connectives, which vary between languages. In addition, automatic word alignment can only be applied successfully if enough parallel data is available, while resources for the language under consideration here are limited.

One possible drawback of using translation text to source connectives is that it draws 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 multiple interpretations, but a translated text might only contain the interpretation of the translator (i.e. in cases where the original connective is disambiguated). Secondly, there are several factors guiding when discourse relations are implicitated or explicitated in translation text, such as specific features of the target language and the relation sense (Becher, 2011; Zufferey and Cartoni, 2014). The distribution of explicit relations in the translation text thus does not necessarily reflect the distribution of explicit relations in the source text. This would influence the likelihood of connectives found in the source language by only taking into account the connectives in the translation text. We will therefore also assess how well the explicit connectives in the translated text reflect the discourse relations in the source text.

Finally, connective lexicons contain information about the relation senses that can be signalled by a certain connective. The PDTB end-to-end parser provides relation sense labels for the connectives it identifies, which could be projected to the connectives in the parallel text. This was, however, not taken into account by Laali and Kosseim (2014). The present research therefore evaluates and extends their methodology for lexicon creation with a parallel corpus.

3 Nigerian Pidgin

Nigerian Pidgin is a contact language spoken all over Nigeria. Although officially a pidgin, it is also often considered a creole (e.g., Courtin et al., 2018), as it is a relatively stable language with a fully-developed vocabulary and grammar. In addition to approximately 5 million native speakers, there are estimated to be almost 100 million speakers of Nigerian Pidgin as a second language. It is often also referred to simply as 'Pidgin' by Nigerian Pidgin speakers themselves.

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).

(1) I say 'toh, me I sabi operate computer small small'.
I said 'well, I know a bit about computers'.

Nigerian Pidgin is also influenced by other European languages (e.g. Portuguese, illustrated by *sabi* in the example above) as well as various indigenous languages it has been in contact with, such as Hausa, Igbo and Yoruba. Nigerian Pidgin is characterized by focus constructions (see e.g. *me* in (1)), which are often indicated with the focus particle *na* (Caron et al., 2019). In addition, serial verb constructions are typical for Nigerian Pidgin. As a result, coordinating conjunctions are less frequent in Nigerian Pidgin than in English (Courtin et al., 2018). An example of such a serial verb construction is provided in Example (2), which has been translated in English with the coordinating conjunction *and*. This suggests that discourse relations might be expressed differently in Nigerian Pidgin.

(2) Con dey hustle, con hustle money go give am.
Then we hustled, got money and went to give it to her.

Linguistic research and resources on Nigerian Pidgin are limited, although this has shifted in recent years. As part of a larger project studying the syntactic and prosodic structure of Nigerian

Pidgin, Caron et al. (2019) are in the process of creating a 500,000 word corpus of spoken data, part of which is already available and will be used in the current study. We also note related work from the NLP community in studying Pidgin to explore new approaches for low-resource natural language generation (Chang et al., 2020) and translation (Ogueji and Ahia, 2019). However, no previous research has examined the discourse structure of Nigerian Pidgin.

4 Methodology

4.1 Data

The data we use in our analysis comes from the gold section of the Naija Treebank (UD NSC Corpus),[1] a parallel corpus of Nigerian Pidgin utterances with English translations. This dataset consists of 140,859 words collected in various locations, and represents 87 speakers. The sampling of speakers aimed at balancing age, sex, education, linguistic and geographic background. The genres recorded cover life stories, speeches, radio programs, free conversations, cooking recipes, comments on current state of affairs, etc. 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).

We extracted the original Nigerian Pidgin utterances (referred to as Source Text – ST), together with their English translations (referred to as Translated Text – TT) and the Nigerian Pidgin POS tags from the NSC corpus. In total, the dataset used in this study contains 9,242 Nigerian Pidgin utterances, divided into three subsets: dev (n=991), train (n=7,279), and test (n=972).

4.2 Construction of the lexicon

Figure 1 displays the workflow used in this study. As Step 1, we ran an automatic end-to-end PDTB classifier on the English text to extract the TT connectives, by disambiguating between connective and non-connective usages of connective tokens. This resulted in the extraction of 4,592 TT connective tokens (dev: n=454, train: n=3,691, test: n=446), consisting of 41 different TT connective types, in 3,389 utterances. In addition, the parser provides the relation sense (PDTB 2.0) of each English connective.

As a next step, the first two authors both manually annotated the Nigerian Pidgin counterpart of all English connectives in the dev set in order to obtain an initial dictionary of English-Nigerian Pidgin connective mappings. Following the criteria from the PDTB (Prasad et al., 2008), only Nigerian Pidgin words from the following syntactic categories were initially considered as connective candidates: coordinating conjunctions, subordinating conjunctions, adverbs, and adpositions. Of 446 TT connectives, 336 were mapped onto an equivalent in the ST. This mapping functioned as input for an initial (seed) connective dictionary, which contained 30 Pidgin connective candidates and 52 unique co-occurrence types (excluding TT connective∼implicit ST pairing).

This initial dictionary was then used to predict the Nigerian Pidgin equivalent of the extracted English connectives in the train set in Step 3. For each utterance containing an English connective in the TT, the POS-tagged ST was searched for one of the Nigerian Pidgin equivalents that had been mapped onto this connective in the dev set. In order to predict a mapping for cases where a single Pidgin connective was mapped to multiple English connectives or vice versa we calculated normalized pointwise mutual information (NPMI) of co-occurrences in our dictionary. The connective match with the highest NPMI was selected as the target connective. For example, English *until* could be a translation of Nigerian Pigin *till* or *sotay*. If the ST of a TT containing *until* would contain both *till* and *sotay*, *sotay* would be selected as its association with *until* is stronger than that of *till*.

In cases where the NPMI did not resolve to a one-to-one mapping (e.g. the TT text contained two instances of *until*, with only one equivalent in the ST), the Nigerian Pidgin connective with the closest relative position in the sentence was selected. When no Pidgin translation equivalent was found, the four words closest to the relative position in the TT were searched for another Pidgin connective that had not been mapped to this English connective yet to extend the initial dictionary with new candidate mappings.

To examine whether the data set contained other Nigerian Pidgin connectives that did not occur in the dev subset, we manually annotated selected instances from the train set where (a) an English

[1] https://universaldependencies.org/treebanks/pcm_nsc/index.html

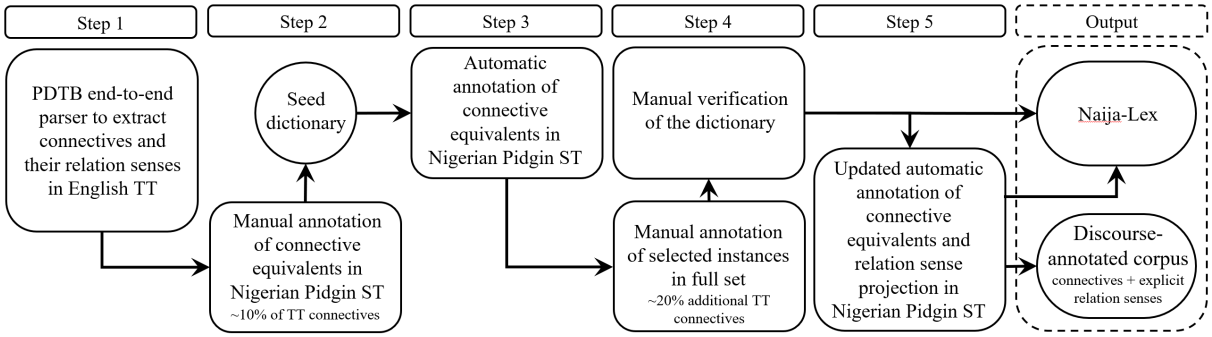

Figure 1: Schematic overview of the workflow for the construction of the connective lexicon (and discourse-annotated corpus) for Nigerian Pidgin.

connective had been identified in the TT that did not appear in the dev set and (b) where no Nigerian Pidgin translation equivalent was found (Step 4). We also manually examined all instances where the Pidgin connective candidates 'wey' or 'sey' were extracted from the ST, since these are frequently used as complementizers rather than connectives. In this annotation round we expanded the connective dictionary by relaxing restrictions on the syntactic category, due to high co-occurrence of English connectives with certain auxiliary verbs and lexical phrases in Pidgin. We then further verified the dictionary by checking the new English-Nigerian Pidgin connective mappings that did not occur in the seed dictionary and consulted a native Pidgin-speaking linguist to verify (a) whether connectives that did not have an English origin were indeed a connective in Pidgin and (b) whether the English translations were valid.

The updated dictionary was used to improve the automatically identified Nigerian Pidgin connectives in the train set (Step 5) and to construct the connective lexicon. The lexicon was complemented with automatic relation sense labels and the frequency data from the corpus.

5 Naija-Lex: A lexicon of Nigerian Pidgin connectives

In total, 4,186 Nigerian Pidgin connective tokens were identified. They were retrieved based on 40 English connective types in the translation text. The Nigerian Pidgin connective tokens consisted of 77 types, which were categorized in 57 entries in the lexicon. Naija-Lex is made available online.[2] 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.

5.1 Connective origin and syntactic category

The majority of the connective entries (n=39) are derived from connectives in the English lexifier. Their phonology has often been adapted slightly to that of Nigerian Pidgin, leading to multiple variants of the same connective. For example, *because* occurs in Nigerian Pidgin as *because*, but also as *cause* and *cos*, and *dough* has been adapted from English *though*. Only 18 connective types in the lexicon did not originate from English connectives. Examples of these are *abi* (English: *or*), *wey be sey* (English: *when*) and *sotay* (English: *until*).

Some of these Pidgin connectives have evolved from English non-connective words. Consider *con* (originating from English 'come') and *make*. Both have evolved from an English verb and can now be used as a main verb in Pidgin, in which case they do not function as connectives, as well as an auxiliary, in which case they are frequently translated as connectives expressing causal and temporal connectives in English (see Examples (3) and (4)).

(3) I **con** realise sey omo na dis Pidgin na im make us connect like dat.
Then *I realized that, wow, this Pidgin brought us together.*

(4) De want hold am **make** e no fall.
They want to hold them ***so*** *they won't fall off.*

Con is used frequently in narration, connecting the events in one utterance with the following events in the next utterance. Like many other

[2] https://osf.io/xns9z/?view_only=710a1eca318f46b9b10584c3b980beec

Nigerian Pidgin auxiliary verbs, *con* is positioned between the subject and the main verb of the clause. By contrast, *make* is only found in clause-initial position. 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. Note that *make* does not satisfy the criterion of not being modifiable (cf. Stede et al., 2018): depending on the pronoun, it is inflected to *meh* or *mey*. We still consider it a connective, because it can also be used to signal a causal relation between two abstract, clausal arguments (similar to causal discourse verbs, see e.g. Danlos, 2006).

Other unique Pidgin connectives have evolved from Pidgin words that originally did not express a discourse function. The connective *naim* is a contraction of the auxiliary verb *na*, which is often used as a focus particle, and the 3SG pronoun *im* and has been lexicalized to a causal/temporal connective, equivalent to *then* or *so*. An example of its connective and non-connective use can be found in (5) and (6) respectively:

(5) E get one lady, hm she just enter inside shop o. **Naim** she say she wan do facials.
There was one lady, she just went into the shop. ***Then*** *she said she wanted to do facials.*

(6) And **na im** dey give am.
And ***he's the one*** *giving it to her.*

The most frequent syntactic category of the connective variants was subordinating conjunctions (n=16), followed by adverbs (n=15) and adpositions (n=11). The lexicon also contains one particle as Pidgin connective (*shey*, English *so*), which is another syntactic class traditionally not considered to contain connectives in English.

5.2 Connective and relation sense distribution

In terms of frequency of connectives in the Nigerian Pidgin data, the data shows that *if* was the most frequent (n=643), followed by *so* (n=606) and *and* (n=511). The most frequent connectives that are not English cognates are *con* and *naim*, with 394 and 77 occurrences respectively.

Table 1 provides information regarding the relation sense distribution found in the corpus, based on projected labels originally assigned to the English connectives. Connectives in Nigerian Pidgin were used to express a total of eight different relation classes (based on PDTB2 level 2 relation labels). Temporal and Cause were the most frequent explicit relations in the corpus. Table 1 also shows that Conjunction relations were frequently implicit in the ST and had been explicitated in the TT. Importantly, the distribution is similar in the manually annotated dev set, indicating that this distribution is not an artefact of the automatic annotation.

Relation sense	Explicit	Implicit
Temporal	30.4	17.8
Cause	25.8	10.3
Conjunction	14.1	52.8
Contrast	11.1	6.4
Condition	15.4	10.6
Alternative	1.4	0.8
Restatement	1.5	1.0
Instantiation	0.3	0.3

Table 1: Relation sense distribution in percentages for connectives that were explicit (n=4049) and implicit (n=388) in Nigerian Pidgin (but explicit in English).

When considering the relational distribution of each connective, we find that most connectives have one clear dominant meaning. This is illustrated in Figure 2. For example, Nigerian Pidgin *den* is almost exclusively used in temporal relations. The connective *as* is predominantly used as a temporal connective, but similar to English, has also maintained its causal meaning. The non-cognate connectives *con* and *naim*, however, are used in various relations, mostly Temporal and Cause. Although it might seem surprising that additional connectives have developed for relations that can also be signalled by English cognate connectives (e.g. *den* for Temporal and *so* for Cause), we note that these relations are also the most frequent, which might explain why speakers are more likely to find new ways to express them. Alternatively, these new connectives might have evolved precisely because of the polyfunctional usage. The connectives borrowed from English clearly signal one dominant relation, whereas *con* and *naim* can be used to underspecify these discourse relations.

6 Evaluation of the method

6.1 Automatic connective identification in TT

To evaluate accuracy of the first step in our pipeline we manually annotated the TT in the development set for presence of connectives, according to the criteria set by Stede et al. (2018).

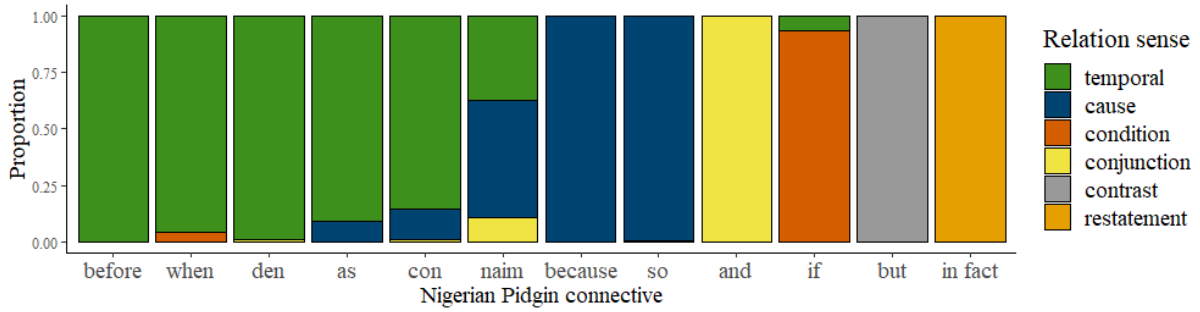

Figure 2: Relation sense distribution of the twelve most frequent Nigerian Pidgin connectives.

Connectives were identified in the English TT by the PDTB end-to-end parser with relatively high reliability (precision = 85.2%, recall = 85.6%, F1 = 85.4%). This is lower than the parser's performance on the test set of the PDTB (i.e. 95.76%), but that was expected, because the parser was trained on text from the written newspaper domain.

A closer examination of the performance per connective shows that precision is especially low on *so, so that, also* and *later* (<= 50%). In spoken language, *so* is often used as a discourse marker rather than a connective.[3] For *so that*, the majority of mistakes were caused by the parser identifying *so* with a demonstrative *that* as a connective. Low recall was reached for *afterwards* and *yet*, which occurred infrequently and were not detected by the parser. The inaccuracy of automatic connective identification is thus mainly due to a small number of connectives. Manual examination of these instances can improve the accuracy significantly.

6.2 Relation sense identification in TT

The PDTB end-to-end parser was also used to extract relation senses for the connectives in the TT. In order to evaluate the parser's reliability, the second author and an independent coder annotated the English connectives in the dev set using PDTB2 labels. Agreement was high (Cohen's κ: 0.88, CI [0.85, 0.92]; AC_1 = 0.89, CI [0.86, 0.92], level 2 labels) (see Spooren and Degand, 2010, for IAA standards in discourse relation annotation). Disagreements were resolved by the first author.

For the connectives that the parser had identified correctly, the label provided by the parser was identical to the gold label in 78.8% of cases (Cohen's κ: 0.74, CI [0.70, 0.79]; AC_1 = 0.77, CI [0.72, 0.81]). Again, reliability was characterized by low performance on a small number of relation senses. More specifically, there was extremely low precision on Concession relations (0.04), as well as low recall on Contrast relations (0.23), because many Concession relations were identified as Contrast relations by the parser. Since the parser cannot distinguish between these relations reliably, these senses were merged in further analyses. Note that these distinctions are notoriously difficult to annotate, even by experts (Robaldo and Miltsakaki, 2014; Zufferey et al., 2012). Similarly, recall was low on Synchrony relations (0.30), since many of them were considered Asynchronous in manual annotation. These senses were therefore also collapsed. Adopting this 8-way classification resulted in very high reliability: (Cohen's κ: 0.90, CI [0.86, 0.94]; AC_1 = 0.91, CI [0.88, 0.95]).

The connective *in fact* was classified by the parser as signalling a Conjunction relation instead of Restatement, leading to a low precision (0.20) on this relation sense as well. The parser classified the two instances of *as if* in the development set as Concession, whereas the manual annotation was Expansion. Note that the PDTB 2.0 manual also indicates that both of these interpretations are possible for these connectives.

6.3 Connective mapping from TT to ST

To evaluate how reliable the automatic connective mapping was with the updated dictionary (i.e. Step 5 in Figure 1) and how the lexicon extends to new data, the accuracy of the final dictionary was determined based on its performance on the test set. 87.9% of the English connectives in the test set were correctly matched with their Pidgin connective. Sixteen connectives in the test set (3.6% of the total TT connectives) were incorrectly identified as implicit in ST, because their Pidgin translation had not co-occurred with this TT connective in the

[3]Note that the difference between discourse marker and connective *so* was reflected in the Pidgin POS-tags. Discourse marker uses of *so* were mostly annotated as adverbs, whereas connective uses of *so* were mostly subordinating conjunctions.

initial set. When allowing for the automatic mapping to identify new TT-ST pairings, the overall accuracy of the mapping in the test set increased to 90.9% (precision: 92.0, recall: 97.9, F1: 94.9).

The remainder of the incorrect mappings were mostly due to the English connective being incorrectly identified by the parser as a connective (n=12) or to one or multiple possible Nigerian Pidgin translations of the English connective occurring in the utterance, while they were not a translation of the English connective. Again, performance on individual connectives was high, except for a few Nigerian Pidgin connectives: *sey, wey, for, na* and *like* all reached F1 scores lower than 70%. This is likely because they also frequently occur with a non-connective meaning.

Finally, we evaluate how many Nigerian Pidgin connectives are missed when annotating only a subset of the corpus manually. In the Pidgin test set, we identified six Pidgin connectives that did not occur in the lexicon so far. Two of them were corpus annotation variants (i.e. different spelling/POS-tag) of connectives that had already been extracted from the initial set. The remainder of the new Pidgin connectives were separate entries and their English TT equivalent had either been mapped onto a new mapping or had incorrectly been identified as implicit in the ST. Extending the lexicon using additional datasets can therefore be reliably done using manual annotation of only those English TT connectives that either have not occurred previously or that are mapped onto a ST connectives that so far has not co-occurred with the English connective.[4]

6.4 Relation sense projection from ST to TT

A final premise of this method is that the relation marked in the TT reflects the relation expressed by the connective in the ST (cf. Laali and Kosseim, 2014). To evaluate this assumption, we manually annotated the relation sense of the Nigerian Pidgin connectives that had been identified in the dev set. Inter-annotator agreement on a subset of the data was sufficient (Cohen's κ: 0.77, CI [0.59, 0.94]; AC_1 = 0.82, CI [0.68, 0.96]). These manual annotations of the ST were then compared with the projected automatic annotations of the PDTB parser. Since the analysis above shows that distinctions within the class of Temporal relations and Comparison are not reliably annotated, these classes were combined. Agreement between the parser's label and the manual annotation on this 8-way classification was high (Cohen's κ: 0.90, CI [0.86, 0.94]; AC_1 = 0.91, CI [0.88, 0.94]) and only slightly lower than the reliability of the parser on the translation text. Using automatic annotation of a translation text to extract the relation sense of connectives in a source text is therefore well feasible.[5]

7 Discussion

The present research shows how extending discourse research to contact languages provides more insight in how discourse structure is realized cross-linguistically. We find a similar relational distribution between English and Pidgin. Moreover, we identified a relatively large variety of discourse markers in Nigerian Pidgin. Taken together, this indicates that the discourse structure of Pidgin is not necessarily less complex than in English, even though pidgins are generally considered to be less syntactically complex.

Furthermore, the analysis reveals that most of the discourse connectives in Nigerian Pidgin are derived from English discourse markers. However, discourse relations in Pidgin are not always marked by lexical items from a limited set of syntactic categories (cf. Stede et al., 2018): Auxiliary verbs can also signal coherence relations in Nigerian Pidgin. We therefore recommend researchers working on identifying connectives in different languages to be flexible in what items might signal coherence relations. Cross-linguistic variation should be taken into account when defining restrictions on what can constitute a discourse connective. In order to do so, more research on coherence marking in different languages (especially non-Indo-European languages) is needed.

Nigerian Pidgin is mainly a spoken language. This affects our analysis in several ways. Firstly, the nature of spoken language influences the distribution of connectives and the relation senses they mark. A smaller number of connective types are used in spoken compared to written text (Crible and Cuenca, 2017). At the same time, connectives in spoken language are used in a wider variety of functions. This should therefore be taken into account when considering the frequency of connectives and their relation sense distribution in the lexicon.

Secondly, the performance of the parser is in-

[4] This amounts to less than 5% of the data for our corpus.

[5] To reduce noise in the corpus and the lexicon, we manually checked the relation senses that accounted for less than 5% of the cases for a given Pidgin connective.

fluenced by the change in domain. A comparison with the parser's performance on a different English discourse-annotated spoken corpus (DiscoSPICE, Rehbein et al., 2016) revealed that the parser makes comparable mistakes in identifying connectives in dialogues by native speakers of English (see also Scholman et al., 2021). Inaccuracies of the parser in the current study are thus for the most part due to the fact that our data is spoken.

Despite the fact that the process described in this paper contains a number of manual steps, we believe that this method can easily be extended to other low-resource languages. The annotations were for the most part done by the second author, who is semi-proficient in Pidgin. In addition, comparing instances of connective candidates that were automatically identified in the corpus allowed a non-native speaker with little knowledge of the language (the first author) to manually verify the data. Although the native speaker we consulted was a trained linguist, the additional required judgments can be also be obtained in such a way that no linguistic knowledge is required from the native speaker. Furthermore, the required amount of manual annotations does not increase considerably with larger datasets. Firstly, by annotating the dev set alone, only a few infrequent connectives had been missed (6 out of 77 connective variants) and annotating a small subset of the data helps to identify these cases. Secondly, larger resources lead to a stronger reliability of statistical approaches. A combination of manual and automatic approaches therefore seems a fruitful way to construct resources for other under-researched languages as well.

Another challenge in identifying connectives in contact languages specifically, is the large amount of variation and the frequent use of code-switching. Especially for an English-lexified language like Nigerian Pidgin, it can be difficult to distinguish between cases in which the utterance is English or Nigerian Pidgin. In fact, some of the connectives that we initially identified as ST connectives were in fact English connectives that had been used during code-switching. Frequency filters, which have also been adopted in fully-automatic approaches to cross-lingual connective identification (Laali and Kosseim, 2014), are therefore crucial. Lexical items that only occur a few times or are only used by a small number of speakers should be checked manually. To distinguish between code-switched and infrequent connective uses, the context should be taken into account. If the connective is embedded in an English-only sentence, it is likely not a connective in the target language.

The similarities between Nigerian Pidgin and English could also have been employed as an advantage. For example, as one reviewer suggests, Levenshtein's distance can be used to find Nigerian Pidgin connective candidates that are similar to the English connective. However, there are a few drawbacks to such an approach. The Nigerian Pidgin version of an English connective sometimes diverged strongly from its English cognate (e.g. *because - cos)* and some of the Nigerian Pidgin connectives with an English origin were not cognates of the English connective, but of a synonym (e.g. the Pidgin meaning-equivalent of *instead* is *rader*). Levenshtein's distance might therefore not always be a reliable measure. More importantly, our goal was to establish the usefulness of existing resources to research discourse structure in low-resource languages in general, not English-lexified contact languages only. In addition, in searching for Nigerian Pidgin connectives, we did not want to create a bias towards English connective cognates, as one of our research goals was to examine how discourse connectives emerge and to establish which connectives would not originate from English.

8 Conclusion

This study presents Naija-Lex, a lexicon of connectives in Nigerian Pidgin, which was constructed by combining automatic discourse parsing with manual annotation. In doing so, we showed that automatic discourse parsing is a valuable resource in constructing a discourse-annotated database to extract connectives. However, construction of a seed set based on a subset of the data and manual annotation of selected connectives and relation senses is necessary to obtain a reliable database. In addition, we made a first step to building a Nigerian-Pidgin discourse-annotated corpus by annotating ST connectives that were explicit in the TT. We will extend this corpus with implicit relations in further work.

Acknowledgements

We are highly grateful to Emeka Felix Onwuegbuzia for his valuable insights on Nigerian Pidgin and evaluating a subset of the connectives. This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

References

Viktor Becher. 2011. When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies*, 23(1):26–47.

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Rome, Italy.

Peter Bourgonje, Jet Hoek, Jacqueline Evers-Vermeul, Gisela Redeker, Ted Sanders, and Manfred Stede. 2018. Constructing a lexicon of Dutch discourse connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *TLT 2019, Treebanks and Linguistic Theories, Syntaxfest*.

Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. 2020. Unsupervised Pidgin text generation by pivoting English data and self-training. In *International Conference on Learning Representations (ICLR)*.

Marine Courtin, Bernard Caron, Kim Gerdes, and Sylvain Kahane. 2018. Establishing a language by annotating a corpus. In *annDH 2018 Annotation in Digital Humanities*, volume 2155, pages 7–11. CEUR.

Ludivine Crible and Maria-Josep Cuenca. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Laurence Danlos. 2006. "discourse verbs" and discourse periphrastic links. In *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache) Universität Konstanz*, page 160.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

Debopam Das, Manfred Stede, Soumya Sankar Ghosh, and Lahari Chatterjee. 2020. DiMLex-Bangla: A lexicon of Bangla discourse connectives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1097–1102.

René Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse (CoDi 2020)*, pages 65–75, Online. Association for Computational Linguistics.

Murathan Kurfalı, Sibel Ozer, Deniz Zeyrek, and Amália Mendes. 2020. Ted-MDB Lexicons: TrEnConnLex, PtEnConnLex. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 148–153.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Amália Mendes and Pierre Lejeune. 2016. LDM-PT: A Portuguese lexicon of discourse markers. In *Conference Handbook of TextLink–Structuring Discourse in Multilingual Europe Second Action Conference*, pages 89–92. Debrecen University Press.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124. Association for Computational Linguistics.

Kelechi Ogueji and Orevaoghene Ahia. 2019. PidginUNMT: Unsupervised neural machine translation from West African Pidgin to English. In *NeurIPS 2019 Workshop on Machine Learning for the Developing World (ML4D)*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

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

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2010, page 657. American Medical Informatics Association.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Livio Robaldo and Eleni Miltsakaki. 2014. Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, 5(1):1–36.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

Merel C.J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the First Workshop on Computational Approaches to Discourse (CoDi 2020)*.

Wilbert P. M. S. Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2018. Connective-Lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Manfred Stede and Carla Umbach. 1998. DiMLex: A lexicon of discourse markers for text generation and understanding. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1238–1242.

Benedikt Szmrecsanyi, Bernd Kortmann, Geoffrey Sampson, David Gil, and Peter Trudgill. 2009. Between simplification and complexification: Non-standard varieties of English around the world. In *Language complexity as an evolving variable*, pages 64–79. Oxford University Press.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the CoNLL-16 shared task*, pages 1–19.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019)*, pages 133–143.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies*, 26(3):361–384.

Sandrine Zufferey, Liesbeth Degand, Andrei Popescu-Belis, and Ted J. M. Sanders. 2012. Empirical validations of multilingual annotation schemes for discourse relations. In *Proceedings of the 8th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-8)*, pages 77–84, Jeju Island, Korea.