Master’s Thesis

Universal Dependencies for Cape Verdean Creole

Universidad del País Vasco | Euskal Herriko Unibertsitatea

Submitted in fulfillment of the requirements for the degree of

Master of Science
in
Language Analysis and Processing

Author: Brandyn Emile Evora
Advisors: Aitziber Atutxa Salazar and Koldo Gojenola

September 9, 2019

DEDICATION

There are many people in my life that have helped me get to where I am now. This thesis is dedicated to my mother, my brothers, my grandmother, our family, my friends, my colleagues at both the University of Massachusetts, Amherst and the University of the Basque Country, and to the people of Cape Verde.

ABSTRACT

The Universal Dependencies Project has been a largely successful attempt to devise an annotation protocol that works cross-linguistically. Following the work done by Ryan McDonald and his team in 2013, a standardized system of annotation has been proposed to allow for more uniform multilingual parsing. The system has been widely adopted and there are currently treebanks for over 80 languages with more to come.

Following the framework laid out by the Stanford Dependencies Treebank for English as well as the part-of-speech tag set created by Google and detailed in Petrov et al. (2012), linguists worldwide have been able to annotate treebanks which allow for cross-linguistic research and application development.

Adhering to the guidelines of the Universal Dependencies Project, I have begun the annotating for Cape Verdean Creole, the oldest creole language still spoken today as well as the most widely spoken Portuguese-based creole.

Cape Verdean Creole, or *kriolu*, is not the official language of the independent archipelago found off the northwestern coast of Africa, yet the several varieties of the creole are used daily by the citizens of Cape Verde as well as by its diaspora found in the United States, Portugal, Angola, France, the Netherlands and many other countries worldwide. With more Cape Verdeans living outside of the country, Cape Verdean Creole is the common link between all of the communities and the culture of the motherland.

The current treebank that I have built contains 528 sentences of the Sotavento variant of the southern islands which were manually tagged for part-of-speech as well for their dependency relations. The sentences were obtained from *Na Boka Noti,* a book of old folk tales written by T.V. da Silva.

DECLARATION

I hereby confirm that the thesis work presented is my own work, and I have acknowledged those who have assisted me.

Brandyn Emile Evora

ACKNOWLEDGEMENTS

I would like to thank my university and its staff for the support I received while carrying out my thesis work as well as my coursework. To the people at the University of the Basque Country in San Sebastian, I am immensely appreciative for the opportunity you provided me to study in your program and for all that I have learned from you about the modern applications of linguistic theory and language analysis practices. I am grateful to my advisors, Aitziber Atutxa Salazar and Koldo Gojenol for their help with carrying out my thesis work.

I am especially grateful for my classmate, roommate and friend, Artur Kulmizev. His friendship and support throughout my coursework and thesis project were immense factors in my success during this program. His support during the training of my models was invaluable.

To my one of my closest friends, Tyler Zentz, thank you for your help fixing the bugs in my code, allowing me to quickly process more text and build a stronger treebank.

I would also like to thank my professors at the university and the members of the IXA group for their willingness to teach my classmates and I the intricacies of the field of natural language processing as well sharing with us the culture and language of the Basque Country.

My thesis would also not be possible without the efforts of the hundreds of linguists and researchers committed to adding to the Universal Dependencies Project. Together we will build a truly impressive database of cross-linguistic information that will hopefully have a large impact on how we develop language applications and carry out our research in the future.

Above all, I would like to thank my mother, Maria Evora, for her constant support and guidance.

TABLE OF CONTENTS

INTRODUCTION

There are thousands of languages that are spoken today throughout the world, and many more that have been lost in history. While languages vary greatly in the sounds the utilize, how they structure words morphologically and in the end how those words are ordered, all languages are theorized to follow the Principle of Compositionality which was formally proposed by Francis Jeffry Pelletier in 1994.

He explains that the semantic meaning of a sentence is derived not only by the meanings of the individual words found in the sentence, but also by the unique order in which they are arranged.

With that in mind, we can deduce that in order to succeed in training artificially intelligent systems which can understand the subtle intricacies of unprocessed human language, we must not only provide databases of word definitions, but also establish an efficient means to which we can process syntax.

Along with the Principle of Compositionality, if we also take into account the work of Noam Chomsky and his theories of language universals, we can begin to devise a strategy to create an annotation scheme that works cross-linguistically.

Chomksy and other linguists have found that there are numerous universal tendencies found between languages that support human language's function of conveying complex meaning effectively and efficiently. In an effort to uncover these universals, languages are analyzed and attempts continue to be made to categorize languages in a meaningful way.

Allowing us to create systems capable of not just following rules built solely around the idiosyncrasies of one language, but able to take in various or all current linguistic information present in written or audible form has always been a goal in linguistics, and the Universal Dependencies Project is a significant step in that direction.

The goal of the Universal Dependencies Project is to cut down on the number of syntactic annotation schemes currently used around the world in order to facilitate research and application development involving all languages currently being studied. Using a standard form of annotation allows us to much more effectively share and compare research findings and apply recently developed processing strategies to new languages.

The end goal of the Universal Dependencies Project is a universal parser that is language independent and allows for the construction of highly functional and accurate multilingual systems. In addition to providing a framework for training artificial intelligence, a universal parser would also allow for linguists to more accurately uncover universal patterns found in languages across the world.

A crucial advantage of utilizing the UD annotation scheme is its ease of application to new languages. Following the formalisms laid out by the leaders of the project, it is very feasible for minority languages and those with little research to be added to the database and contribute to the advancement of the project.

With this thesis I aim to add Cape Verdean Creole to the list of languages represented by the Universal Dependencies Project. Cape Verdean Creole, being the oldest creole spoken today, has a rich history and speakers can be found in many parts of the world.

Creole languages are notoriously difficult to study because being an amalgamation of several languages, they contain vocabulary and syntactic rules adopted from the original language influences as well as novel words and phrases developed by the native speakers of the creole.

Cape Verdean Creole has spread around the world thanks to the large groups of the country’s diaspora found in Europe, Africa and the United States who continue to speak the language in their daily lives. The country’s famous musicians have also been significant contributors in sharing Cape Verdean Creole around the world.

There are several challenges associated with the study of Cape Verdean Creole. The simple nature of the country provides a huge obstacle in confidently defining the language. Cape Verde is an archipelago of 10 islands, 9 of them being inhabited. Within the language there are two distinct dialects, that of the northern islands, Barlavento, and that of the southern islands, Sotavento. Along with having a northern and southern distinction, there are also differences in vocabulary and pronunciation amongst the individual islands.

As a former Portuguese colony, the official language of the country is Portuguese, and as of yet, Creole has not been declared an official language even though it is the native language of virtually all of its inhabitants. Creole is the language of the day to day life of the people, whereas Portuguese is largely reserved for academic, political and formal situations.

In order to create a largely uniform treebank, I decided to focus my work on the Sotavento dialect, as it is the one spoken by my family and therefore the more practical choice for me in terms of understanding the intricacies of the language's syntax. I also used the first official Cape Verdean Creole to English Dictionary written by Manuel Da Luz Gonçalves in order to attempt and assure that the text I used followed the ALUPEC style of spelling.

UNIVERSAL DEPENDENCIES

The Universal Dependencies Project started out as an effort to create a language independent annotation scheme that would be able to be applied to all languages regardless of their unique properties such as agglutinative case markling, lack of copular verbs or pronoun dropping.

Combining the formalisms of the Stanford Dependencies (de Marneffe et al., 2014), Google's Dependencies treebank (MacDonald et al., 2013), Google's part-of-speech tag set (Petrov et al., 2012) as well as the Interset morphosyntactic tagset used in the HamleDT treebank project (Zeman 2012), the Universal Dependencies project was born as an attempt to unify the efforts of these works with the goal of creating one standard annotation practice amongst linguists.

After several attempts to create a standard for syntactic annotation the Universal Dependencies Project finally gained traction in October of 2014. It utilized components of various tag sets, as well as the CoNLL-X format which was later revised to its current form, CoNLL-U.

Using the tools provided by the Universal Dependencies Project we are able to not only annotate the syntactic relationships between words of a sentence, but we are also able to effectively mark morphological features at the word level which in the end results in a richer semantic representation.

While having a detailed representation of the intricacies of semantic meaning is the goal of annotation and parsing, the UD Project must balance out a rich inventory of tags/relations with its overall goal of being language independent. That is to say that as the annotation scheme becomes more precise, it must also remain flexible enough to be applied to language's of different typologies and we have to avoid adding features that only apply to a handful of languages. Ideally we are annotating features that are found cross-linguistically and only adding language-specific tags and relations when absolutely necessary.

As the UD Project is updated and refined, it should in theory, more closely resemble a universal grammar, allowing us to parse any and all languages with a high level of accuracy. The large success of the UD Project can be linked to how quickly it grew. After the annotation guidelines were devised in October of 2014, within the first year of its creation, there were three releases of treebanks. The 10 initial language treebanks were added in January of 2015. Followed by a subsequent release of 18 more language treebanks in May of 2015, and in November of the same year, 33 more language treebanks were added to the project.

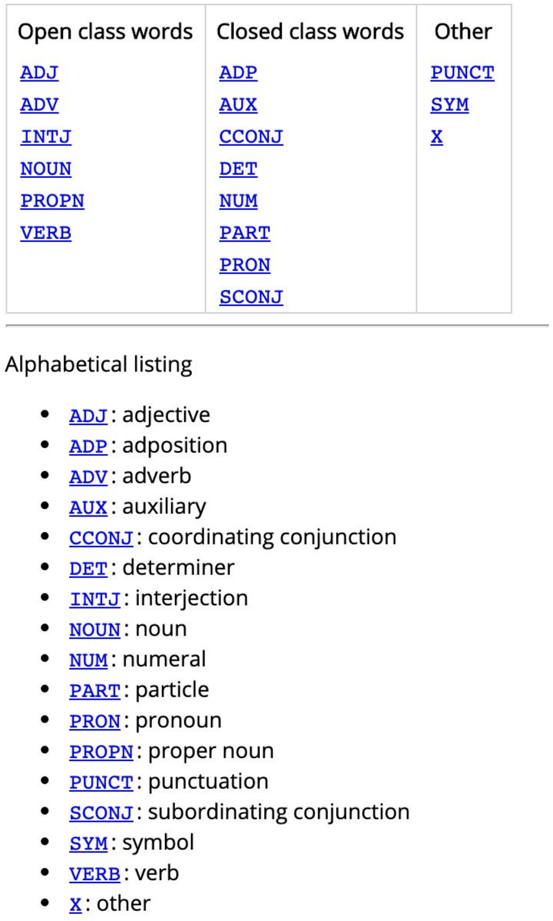
The quick acceptance of the annotation scheme allowed for a large amount of cross-linguistic data to be analyzed and used to further refine the UD Project. Researchers were able to see how well the tags and relationships worked with languages of various typologies as well as where the guidelines would need to be adjusted in order to accommodate language-specific phenomena.

UNIVERSAL DEPENDENCY FORMALISMS

The base unit for UD is the syntactic word. Contractions found in languages like English and Spanish, while orthographically combined, are subject to deconstruction so that the individual syntactic components can be annotated effectively.

However the CoNLL-U format is devised in such a way that the original contracted form is still retained in the annotation with the annotations of the components following it. Multi-word expressions and compounds words have unique ways of being annotated as well, as their syntactic meanings are heavily reliant on the sum of their parts.

Words are marked for their part-of-speech, with 17 different tags currently available. The part-of-speech tags are categorized in three groups: open class words, closed class words, and others.



Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Alphabetical listing

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

Open class words are words that are nouns, verbs, adjectives and adverbs. The simplest way to think of open class words is to think of them as words that can always be created in a language. As languages evolve, new things are invented and new concepts are discussed, and because of that new words need to be created to express them. Think of words like “tweeting” and “laptop”.

On the other hand, closed class words are function words that once established tend to not change often, and it is very rare to add a new one to a language. Modal words like “will” and “must” are great examples, as well as prepositions like “in”, “under” and “beneath”. The category for the others refers to symbols and punctuation.

The UD formalism also allows for the addition of features that may or may not be found in all languages. All language treebanks added to the project must also be submitted with unique documentation outlining any additional modifications to relations or tags found in that specific language that are required in order to accurately annotate the language’s syntax. This allows us to shape the annotation scheme in such a way that it can be tailored to a certain language but also adhere to the guidelines of the UD Project.

There are currently 49 different features that can be used to further enrich lexical annotations. Features include things like nominal gender, verbal agreement of gender and/or number, politeness, definiteness, case, mood, tense, reflexivity, and voice.

- `Abbr`: abbreviation
- `AbsErgDatNumber`: number agreement with absolutive/ergative/dative argument
- `AbsErgDatPerson`: person agreement with absolutive/ergative/dative argument
- `AbsErgDatPolite`: politeness agreement with absolutive/ergative/dative argument
- `AdpType`: adposition type
- `AdvType`: adverb type
- `Animacy`: animacy
- `Aspect`: aspect
- `Case`: case
- `Clusivity`: clusivity
- `ConjType`: conjunction type
- `Definite`: definiteness or state
- `Degree`: degree of comparison
- `Echo`: is this an echo word or a reduplicative?
- `ErgDatGender`: gender agreement with ergative/dative argument
- `Evident`: evidentiality
- `Foreign`: is this a foreign word?
- `Gender`: gender
- `Hyph`: hyphenated compound or part of it
- `Mood`: mood
- `NameType`: type of named entity
- `NounClass`: noun class
- `NounType`: noun type
- `NumForm`: numeral form
- `NumType`: numeral type
- `NumValue`: numeric value
- `Number`: number
- `PartType`: particle type
- `Person`: person
- `Polarity`: polarity
- `Polite`: politeness
- `Poss`: possessive
- `PossGender`: possessor’s gender
- `PossNumber`: possessor’s number
- `PossPerson`: possessor’s person
- `PossedNumber`: possessed object’s number
- `Prefix`: Word functions as a prefix in a compund construction
- `PrepCase`: case form sensitive to prepositions
- `PronType`: pronominal type
- `PunctSide`: which side of paired punctuation is this?
- `PunctType`: punctuation type
- `Reflex`: reflexive
- `Style`: style or sublanguage to which this word form belongs
- `Subcat`: subcategorization
- `Tense`: tense
- `Typo`: is this a misspelled word?
- `VerbForm`: form of verb or deverbative
- `VerbType`: verb type
- `Voice`: voice

Moving past the lexical level, we begin to annotate the relations between words and the UD Project's second version outlines 37 distinct dependency relations which are subject to further specification depending on the language. Relations may be amended to express the use of passive nominal or clausal subjects. We are also able to distinguish between finite clauses capable of standing alone versus non-finite clauses that are dependent on another word in the sentence using the *ccomp/xcomp* relations.

Universal Dependency relations are dependent on the relations between content words; nominals and clauses. Function words such as prepositions and conjunctions are secondary in terms of semantic meaning and only serve as markers of case in many instances. Focusing more heavily on content words also allows for more transfer between one language and another.

Often times the use of function words in one language does not directly correlate to another language's use. We can see examples where one preposition in a given language can be used in multiple syntactic structures where comparable syntactic structures in a different language may use multiple prepositions. In these instances there isn't a parallel relationship between the function words. However we very often see that it is fairly simple to translate one content word from one language to another.

That being said, focusing on the relations between content words allows for us to more consistently annotate and represent semantic meaning cross-linguistically under one annotation scheme.

Following the UD guidelines every word in a sentence is described as being the root of the sentence of a dependent of the root or another word with all words tying back to the root at some point.

RELATED WORKS AND THEIR CHALLENGES

BASQUE

When looking at the efforts of other researchers' to add languages to the Universal Dependencies Project, the first that came to mind was the treebank built by both of my advisors as well as the IXA research group at the University of the Basque Country. Their work, to convert the already existing treebank for Basque to one that followed the guidelines laid out by the UD Project, resulted in the successful automatic conversion of a large percentage of the original data.

Building a treebank based on the UD guidelines for Basque proved to be challenging due to the nature of the language itself. Basque is an agglutinative language which means that complex words are built by the addition of prefixes, suffixes and infixes. It also has a free word order which relies on the use of fourteen cases to provide speakers with information on the syntactic and semantic roles of each word. With an immense potential to create unique words using inflection, one can imagine that training a statistical model would be very difficult.

A hurdle the team faced was lining up their original part-of-speech tag set with the set used by the Universal Dependencies Project. In the case of verbs, where the original treebank differentiated between two classes of verbs, those in need of an auxiliary and those able to stand alone, the team had to forgo this distinction due to the UD tag set's limitation to just one tag for verbs. In order to retain this linguistic annotation, the researchers were required to incorporate the distinction when describing morphosyntactic features of each verb.

The researchers also had to deal with the presence of agglutinative multi-words such as compound nouns and complex postpositions. At the time of publishing their paper on the treebank conversion, they had accounted for roughly two-thirds of these multi-words.

In the end, the team succeeded in developing automatic conversion rules to bring their treebank up to par with the majority of the guidelines of the Universal Dependencies Project.

PORTUGUESE

In the paper by Rademaker et al. (2017) they describe their work converting the pre-existing Portuguese corpus, Bosque, into a treebank following the UD formalism using the parser PALAVRAS. The team updated the original Portuguese treebank with respect to the new UD version.

The use of the Bosque corpus saved a lot of time, because it had already been annotated for dependencies, and so their task was only to devise an algorithm to update the annotations where it was necessary. As they describe, the changing of tag sets from one formalism to another proved to be much less straightforward than anticipated as it was a more subjective task in many instances. This made it difficult for simple conversion rules to be applied.

In the end the team had to coordinate manual revisions with an automatic parser. Interestingly enough, they chose not to work with a form of the corpus which was already in the CoNLL-U format. They argued that using it would have resulted in a loss of linguistic information. They also mentioned that the original CoNLL-U formatted corpus at the commencement of their work was not annotated exclusively by native speakers and they felt that this would detract from the accuracy of the work.

They went on to outline the improvements that they were able to add. This included the retention of complex verb tense information as well as differentiating between multi-word verbal expressions and those merely being modified nominally with an attached preposition. Their main challenges that they described were dealing with contractions and multiword expressions.

IRISH

Much like the group of researchers working with Portuguese, the creation of the Irish UD Treebank was a result of converting a corpus that had already been annotated for dependencies. In 2016, Teresa Lynn and Jennifer Foster from Dublin City University, published their work on the mapping of the Irish Dependency Treebank to the updated formalisms of the UD Project.

In the end, due to the nature of the Irish language, only twenty-six of the universal dependencies were used along with nine language subordinate labels. While a lot of the mapping of tags and relationships was done automatically, there were several instances where relationships in the original Irish Treebank had multiple corresponding relationships in the UD Project's guidelines. In these cases manual correction was required. This could be seen with the use of a *quant* label which could be interpreted as either a *nummod*, *list* or *advmod* in the current version of UD.

Significant changes had to be made manually in terms of the head for coordinations and subordinate clauses. Originally the coordinating conjunction was seen as the head, whereas in the current UD scheme, the head is the first component of the coordination. With respect to subordinate clauses, the subordinating clause was dependent on the conjunction which was in turn dependent on the matrix verb. We see that this has been flipped in the UD guidelines. The clause is directly dependent on the matrix verb with the conjunction dependent on the subordinate clause.

While their work has yet to be finished at the time of publishing, the conversion of the Irish Treebank outlined many challenges that have been faced cross-linguistically.

HIGHLIGHTS OF RELATED WORKS

In reading the publications of researchers who took on the task of converting current dependency treebanks in order to have them follow the established guidelines, it was evident how important it is for linguistics to agree on a standard of annotation that can be applied across all languages.

The unique characteristics of languages that were the leading causes of having so much variation in dependency annotation are important avenues to explore. Having one system of annotation allows us to analyze and compare the performance of linguistic applications on various languages in a much more meaningful way.

Along with the conversion of pre-existing treebanks, continuing to add new treebanks from raw, hand-annotated texts is an important step in cross-linguistic research.

CAPE VERDEAN CREOLE: HISTORY AND LINGUISTICS

Cape Verdean Creole (CVC), or *kriolu* as it is referred to by its speakers, is a creole language spoken in the Republic of Cape Verde and it is spoken by virtually all of its inhabitants as well as by those living in the country's diaspora.

The creole can actually be divided into two distinct dialects, the Barlavento and Sotavento dialects, which correspond to the northern and southern islands. While a speaker of one can understand a speaker of another, the two are easily distinguished by those familiar with the language.

As of 2017 there were estimated to be about 871,000 speakers of CVC, and a large number of these speakers are found outside of the country.

There are large groups of Cape Verdean descendents in Portugal, Angola, the Netherlands, France, São Tome and Principe, Spain, Luxembourg and the United States (predominantly in the New England area).

While CVC is the native language of almost all of those living in the archipelago, the language has yet to be recognized as an official language for the republic.

Prior to July 5, 1975, Cape Verde was a colony of Portugal and therefore all academic, political and formal communication was carried out in Portuguese. The use of Portuguese remains to be the common practice in these situations, resulting in most Cape Verdeans speaking CVC as well as Portuguese in their daily lives.

Cape Verdean Creole is an interesting language to study because by the very nature of being a creole it is subject to the unique characteristics of many languages being combined over the course of the land's history. In 1456 when the Cape Verde islands were discovered by Portuguese travellers they began to cultivate and colonize the islands with the use of slaves taken from Western Africa.

Creole languages are often looked down upon because they tend to originate from slave populations, and while CVC is the world's oldest creole that is still spoken, there is still much research to be done linguistically.

Modern Cape Verdean Creole takes most of its lexicon and phonetic influence from colonial-era Portuguese as well as from the African languages spoken by the slaves brought over. The language also has influence from languages spoken by other immigrant groups such as French, English and other European languages.

Due to the use of Portuguese in schools and in most literature, Cape Verdean Creole has been written and spoken in very different ways depending on the island and region the speaker or author is from. Because of the favoring of Portuguese over CVC in writing, it is hard to find large bodies of text that maintain a uniform style of spelling and grammar. Efforts have been made to standardize the language, but progress has been slow.

In 2005, 30 years after the country's successful secession from Portugal, the Republic's government finally recognized the *Alfabeto Unificado para a Escrita do Caboverdiano* (ALUPEC), an alphabet and spelling paradigm which has helped stabilized the written language. However, this writing system is neither official nor mandatory. ALUPEC utilizes characters based on Latin script and has 28 letters including 5 diagraphs.

Alfabeto Unificado para a Escrita do Caboverdiano

A B D DJ E F G H I J K L LH M N NH O P R RR S T TX U V X Y Z

An immense amount of literature and text written in CVC does not follow this alphabet however making it difficult for CVC to be incorporated in modern linguistic research which uses large bodies of texts to carry out model training. The large disparity in which Cape Verdean Creole is written was a large challenge for the carrying out of this thesis.

Cape Verdean Creole's closest linguistic relative would be Portuguese as it is the most commonly spoken Portuguese-based creole in the world. As such, most of the lexicon comes from Portuguese, yet it's syntactic structure takes heavy influence from other languages found in Africa and Europe.

The largest difference between CVC and Portuguese is in the verbal morphology. Unlike Portuguese, there is very little verbal inflection aside from the imperfect past tense and participle forms of verbs. In Portuguese we see that the root verb of a sentence must be conjugated to agree with the subject in terms of person and plurality as well as with respect to tense, aspect and mood.

This is largely abandoned in CVC, where there is a heavy reliance on auxiliary verbs to convey tense as well as aspect. There is also a large reliance on context that is used to derive the tense and mood of a verb. Oddly enough when a root verb is unmodified in CVC, much unlike languages like Portuguse, Spanish or English, the default tense of the verb is not the present but the past (past perfect to be more specific). Non-copular verbs cannot be expressed in the present or future tense without the use of auxiliary verbs.

In terms of word order, negation, and the expression of possession, we begin to see much more similarity with Portuguese. However, a subtle difference between Portuguese and CVC is the almost total lack of definite articles. One does not say something like, “the woman” in CVC, instead speakers will say something like, “that woman there” or “that woman here”.

CVC also forgoes most use of grammatical gender and we tend to only see gender agreement between certain adjectives and indefinite articles when referring to animate objects.

CAPE VERDEAN CREOLE: ANNOTATION GUIDELINES

In order to build a treebank following the guidelines of the Universal Dependencies Project, I looked towards finding a large enough body was uniform Cape Verdean Creole text. Utilizing samples of texts from numerous sources proved to be challenging due to a lack of uniformity regarding spelling. Having several forms of spelling for the same word leads to issues in terms of training models for parsers.

That being the case, I decided to only use text that I acquired from, *Na Boka Noti* by author T.V. da Silva. The book consists of short stories from Cape Verdean folklore, and although the author used different ways of spelling the same word on several occasions, the text in the book conformed to the ALUPEC alphabet and corresponded to the vast majority of the spelling used in the official Cape Verdean Creole to English dictionary by Manuel Da Luz Gonçalves.

His dictionary was a very important resource during the annotation work of this thesis and coincidentally enough, my mother and the author, Manuel, are good friends and he held his book launch event in my backyard a couple of years ago.

Tokenizing CVC is very similar to languages like Portuguese and English. Words are separated by spaces, with contractions being formed with the use of apostrophes and hyphens.

PART OF SPEECH TAGGING

In terms of part-of-speech tagging, I was able to effectively tag all of the syntactic words using those provided in the standard UD tag set.

1. **Adjectives-** Used to modify nouns, adjectives can be found preceding and succeeding nouns in CVC.
   Examples:
   - **pikénu - "little"**
   - **kunpridu - "long"**
   - **nóbu - "new"**

2. **Adpositions-** Just like we find in Portuguese, CVC uses prepositions to precede nouns and add case or a marker.
   Examples:
   - **ku - "with"**
   - **na - "in"**
   - **sen - "without"**

3. **Adverbs-** Used to modify verbs and adjectives, adverbs can be found both before and after either a verb or an adjective. They are also seen occurring between auxiliary verbs and the main verb of the sentence.
   Examples:
   - **sénpri - "always"**
   - **dja - "already"**
   - **gósi - "now"**

4. **Auxiliary Verbs-** Used to convey tense and aspect, often used with another auxiliary verb, they precede the main verb.
   Examples:
   - **ta - Used to express the future tense as well as habitual actions in the past or present**
   - **sa - Used to express a progressive action in the past or present**

5. **Coordinating Conjunctions-** They are used to link two components of the same type; whether that be nominals, clauses or modifiers.
   Examples:
   - **más - "but"**
   - **y - "and"**

- **o - “or”**

6. **Determiners-** Although there is no parallel to the word “the” as we would find in English or “o/a/os/as” as we find in Portuguese, CVC uses determiners prior to definite/indefinite nouns as well as to take the role of a pronoun in some cases. (Occasionally definite articles are used in multi word expressions taken directly from Portuguese).
   Examples:
   - **un/uma - “a”**
   - **si - “his/her/their”**
   - **nha - “my”**
   - **bu - “your”**
   - **tudu - “all”**
   - **kel/kes - “That/Those”**

7. **Interjections-** Interjections are used to express emotions or opinions and are typically found in discourse. In my dataset the only interjection found is the word “nau/na” simply meaning, “no”.
   Examples:
   - **nau - “no”**

8. **Nouns-** Nouns can be defined as a person, place or thing. In CVC they can inflected to express number and gender.
   Examples:
   - **nóti - “night”**
   - **bizinha - “neighbor”**
   - **rapás - “guy”**
   - **mudjer - “woman”**
   - **makáku - “monkey”**

9. **Numerals-** Whether it digital or written form, numerals modify nouns to express a specific quantity. They can act as determiners or pronouns.
   Examples:
   - **dos - “two”**
   - **sinku - “five”**

10. **Particles-** Particles are used to add meaning to other words. In the case of this dataset, the only particle is the word, “ka”, which is used to negate verbs.
    Examples:
    - **ka - Verbal Negation**

11. **Pronouns-** Taking the place of nouns, pronouns act similarly and can be found as the subject, object, indirect object of a verb or as a nominal modifier. They reference nouns, and their referent can be deduced from context.
    Examples:
    - **N - "I"**
    - **nu - "You"**
    - **el/e' - "He/She"**
    - **es - "They"**
    - **nu - "We"**

12. **Proper Nouns-** Refer to the names of specific people, places or things. Much like Portuguese and other European languages, the first letter of each word in a proper noun is capitalized.
    Examples:
    - **Lisbôa**

13. **Punctuation-** These are symbols used to convey linguistic information such as being a question. The same punctuation found in Portuguese is used in CVC with the additional usage of apostrophes as ways to convey contractions much like we find in English.
    Examples:
    - **. - period**
    - **? - question mark**
    - **! - exclamation mark**

14. **Subordinating Conjunctions-** In CVC, these are used to mark subordinate clauses which are complements of to other clauses.
    Examples:
    - **ki - "That/Who"**
    - **komu - "like"**

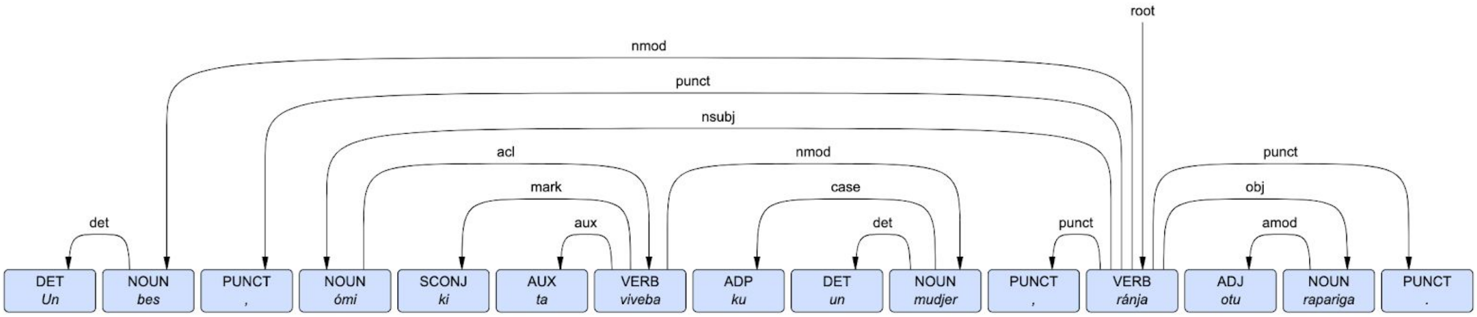
15. **Symbols-** Characters such as currency symbols and mathematical symbols.
16. **Verbs-** A class of words used to denote an event or action. They are inflected to express tense, aspect and mood and are the minimal component for a clause. Participle forms of verbs can act as adjectives and gerundive forms may act nominally.
    Examples:
    - **ten - "to have"**
    - **skrebe - "to write"**
    - **fla - "to say"**
    - **kudi - "to answer/respond"**

- **ba/bai - “to go”**
- **txiga - “to arrive”**

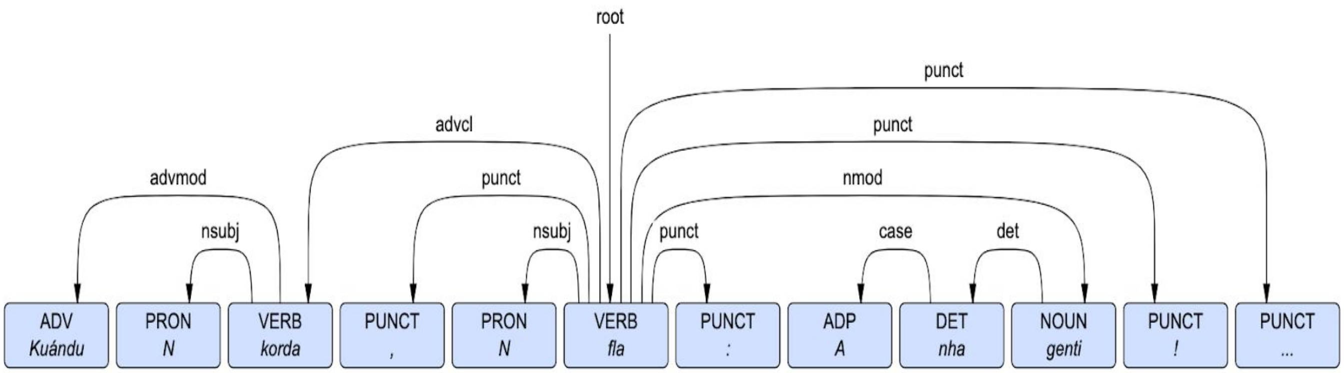
17. **Other -** This class can refer to a number of words which for some reason cannot be assigned a part-of-speech tag such as in the case of the data I annotated, “kó-kó-kó-kó”, the attempt to write out the sound made by a rooster.

DEPENDENCY RELATIONS

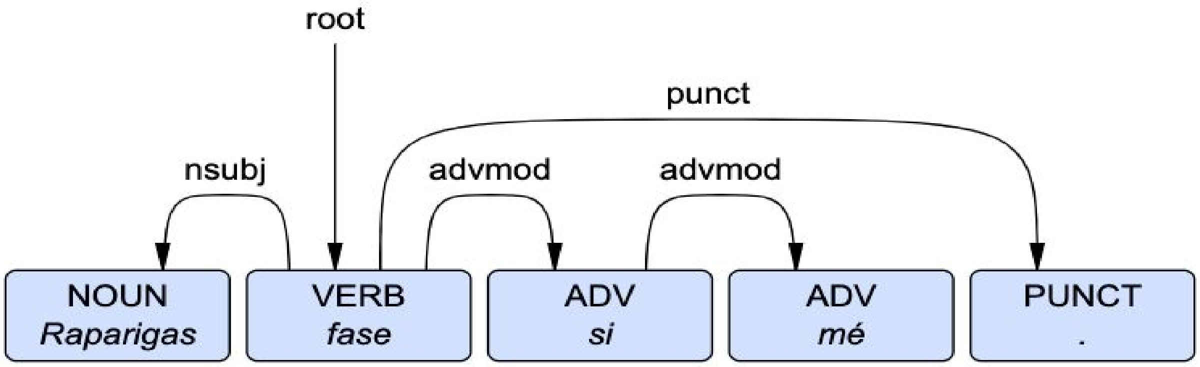
1. **Adjectival Clausal Modifier (acl)-** A clausal modifier of nominals.



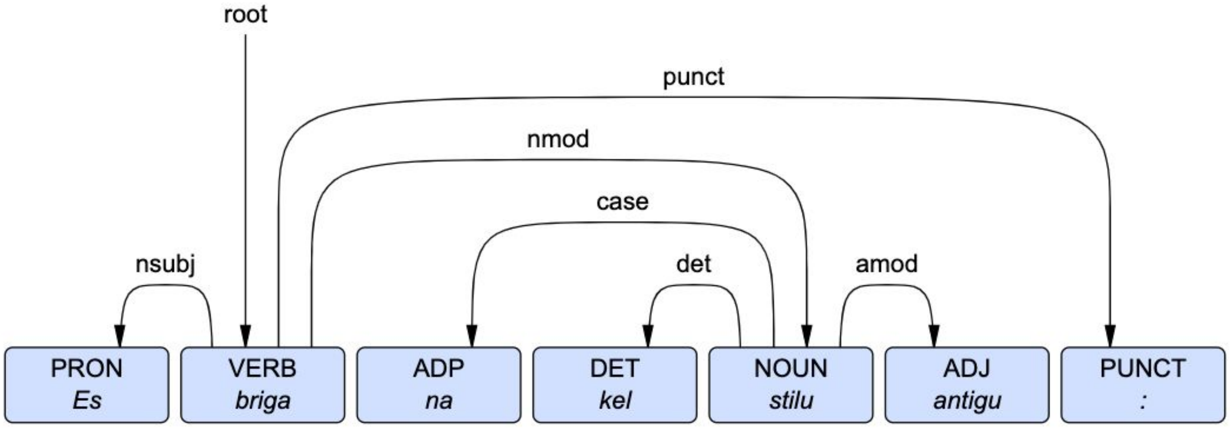
2. **Adverbial Clausal Modifier (advcl) -** A clausal modifier of verbs.



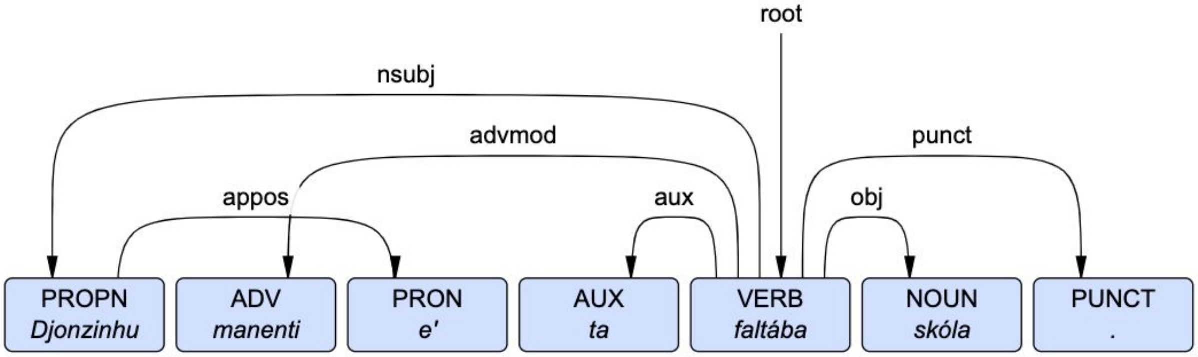
3. **Adverbial Modifier (advmod) -** An adverb that modifies verbs, adjectives or other adverbs.



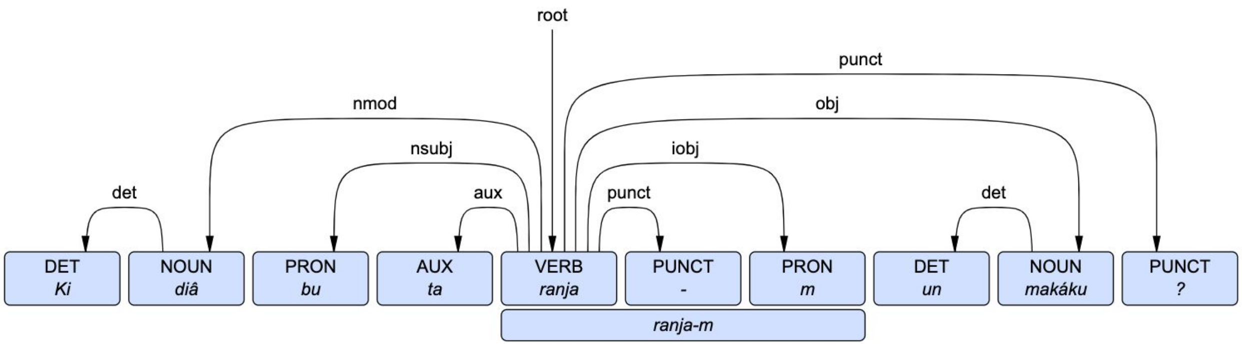
4. **Adjectival Modifier (amod) -** An adjective used to modify nominals. They also act as complements to some verbs such as *fika* ("to stay").



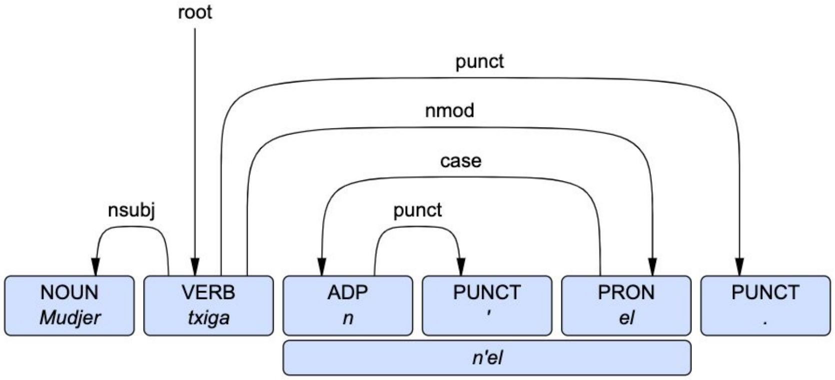
5. **Appositional Modifier (appos) -** A nominal modifier that is adjacent to another nominal which it modifies in order to define or describe it.



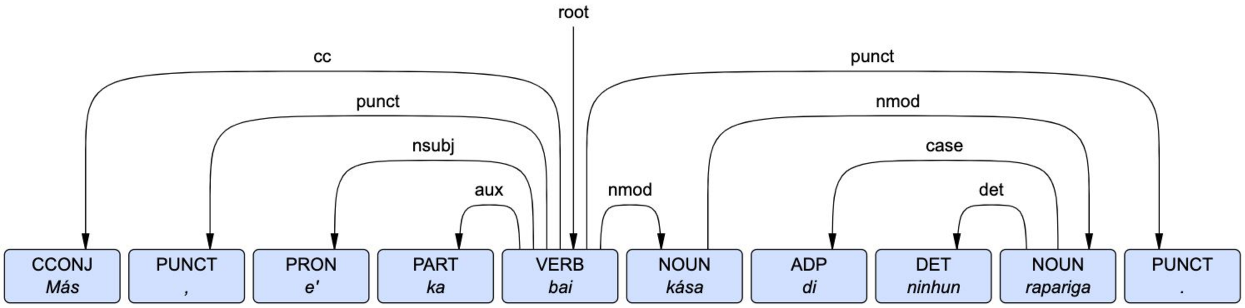
6. **Auxiliary (aux) -** A function word that modifies a clause to express tense, aspect, mood, voice or evidentiality. Clauses in Cape Verdean Creole are also used to express habitual actions.



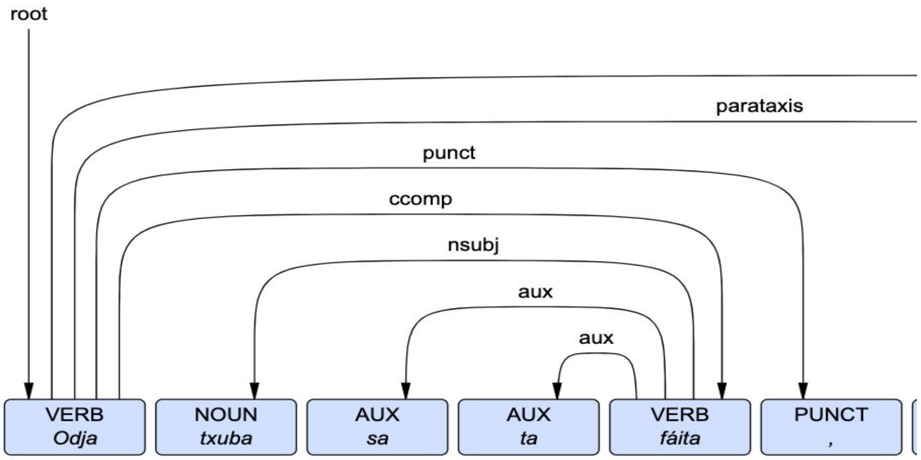
7. **Case (case) -** Dependents of the nominals they attach to or introduce, case markers are often adpositions or clitics.



8. **Coordinating Conjunction (cc) -** A relational marker between two or more conjuncts. It is a dependent of the first conjunct.

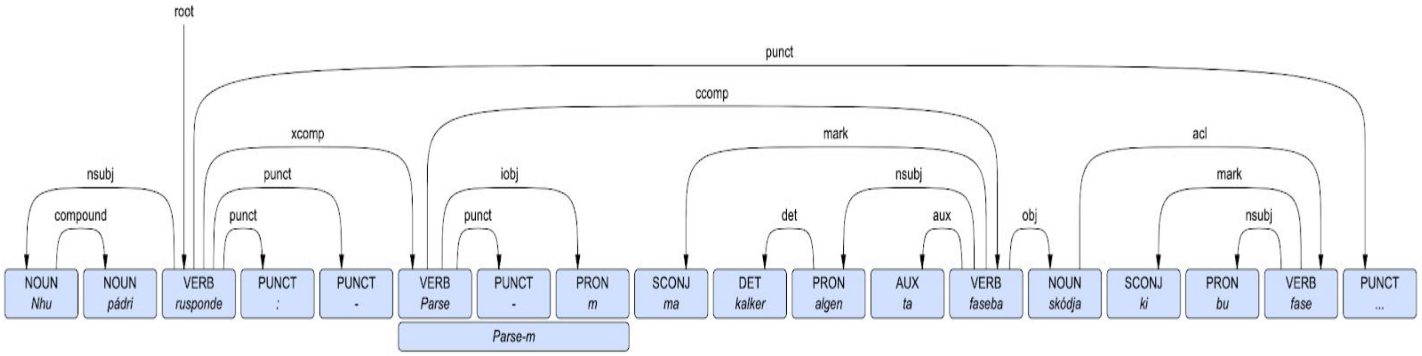


9. **Clausal Complement (ccomp) -** A dependent clause of a verb or adjective which is a core argument.

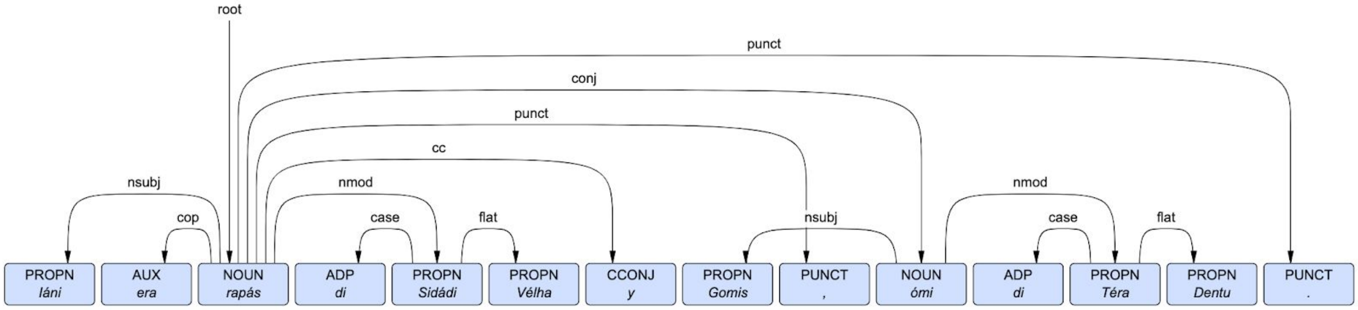


10. **Classifier (clf) -** A nominally dependent used in certain grammatical contexts. Languages that use classifiers are typically found in Asia.

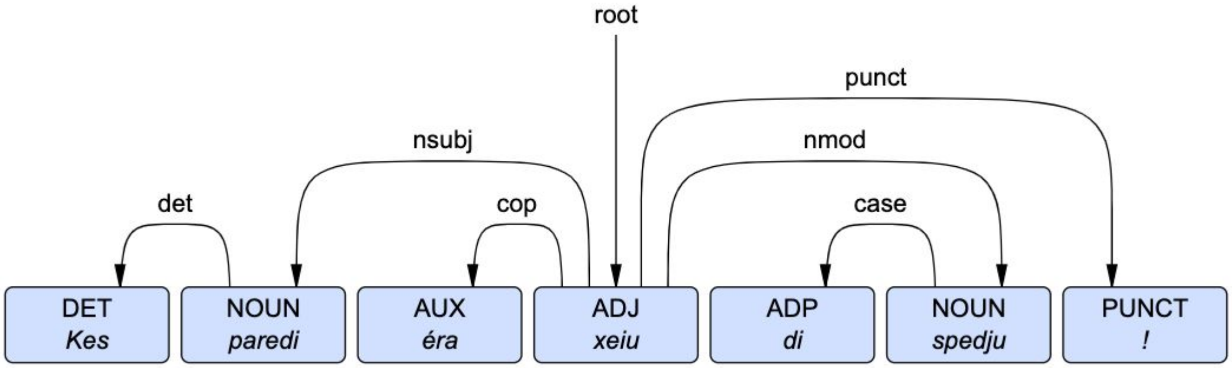
11. **Compound (compound) -** A relationship used to describe compound nouns as well as verbs and adjectives.



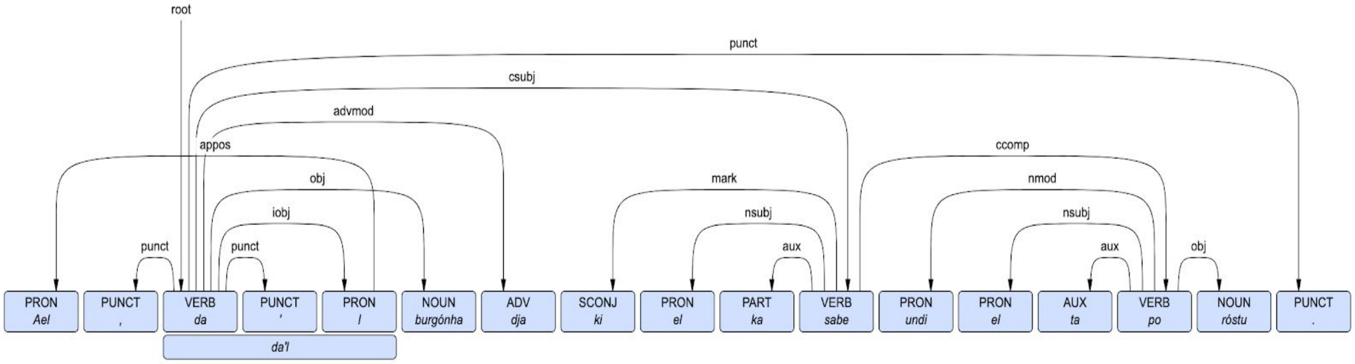
12. **Conjunct (conj) -** The relationship between two or more conjuncts connected by a coordinating conjunction. The first conjunct is the head and the others are connected to it via the conjunct relation.



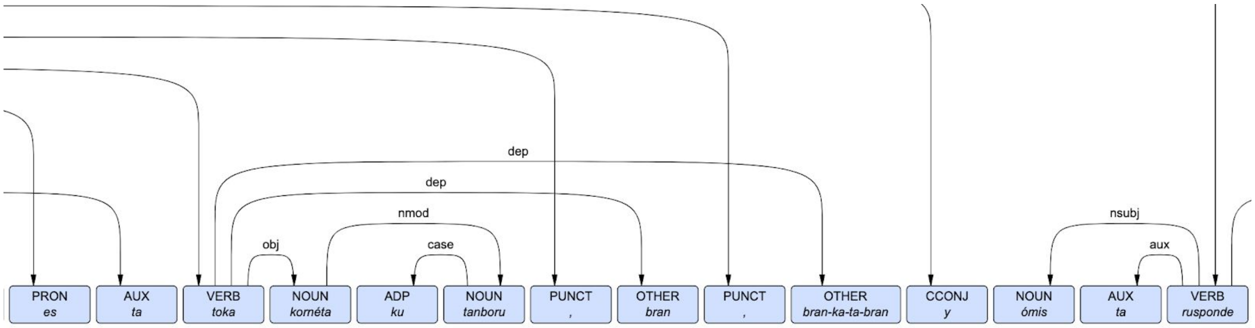
13. **Copula (cop) -** A relation between a function word used to link a subject to a non-verbal predicate.



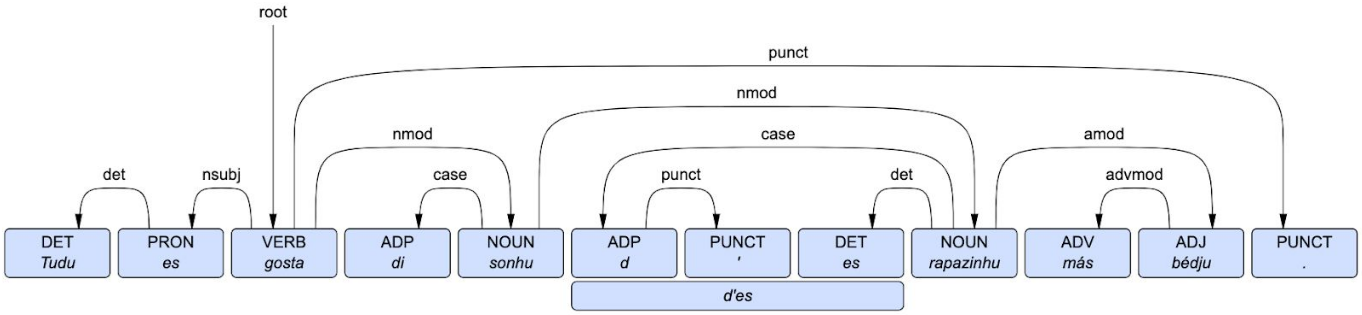
14. **Clausal Subject (csubj) -** The clausal syntactic subject of a clause, that is to say that the subject of the root verb is a clause itself.



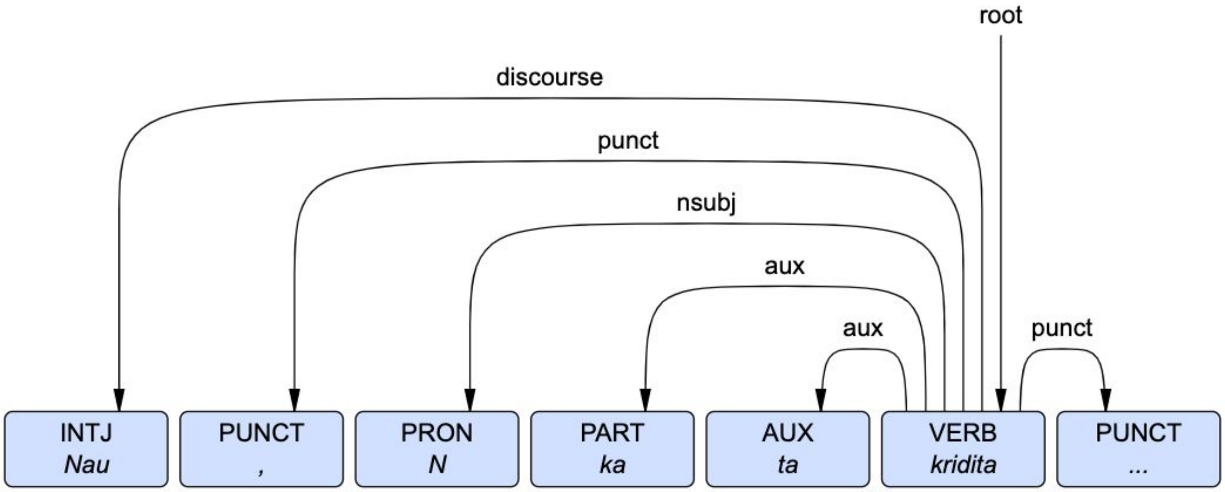
15. **Unspecified Dependency (dep) -** Used to describe a relationship that is impossible to describe more precisely.



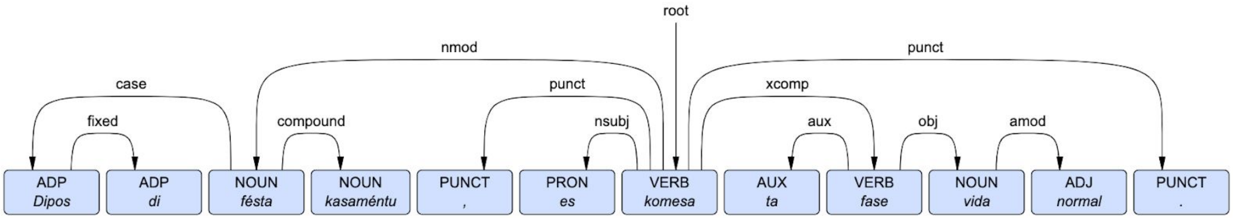
16. **Determiner (det) -** Describes the relationship between a nominal head and its determiner.



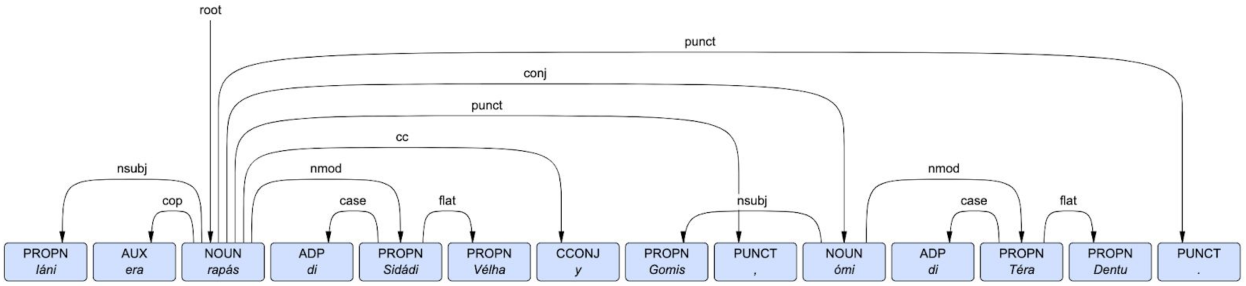
17. **Discourse Element (discourse) -** Used to mark interjections and other discourse elements.



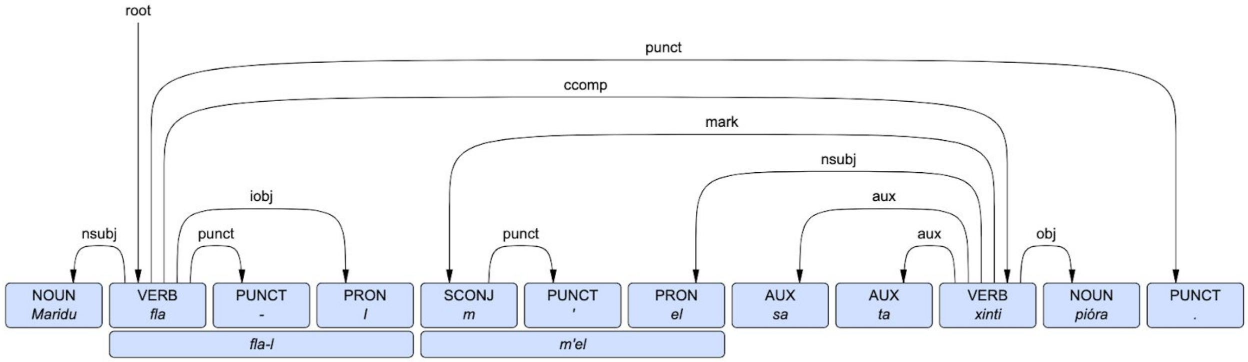
18. **Dislocated Elements (dislocated) -** Used for fronted or postposed elements that usually do not fulfill the core grammatical relations of a sentence.
19. **Expletive (expl) -** This relation captures expletive or pleonastic nominals. These are nominals that appear in an argument position of a predicate but which do not themselves satisfy any of the semantic roles of the predicate.
20. **Fixed Multiword Expression (fixed) -** Used to describe multi-word grammaticized expressions that act as adverbial phrases or function words.



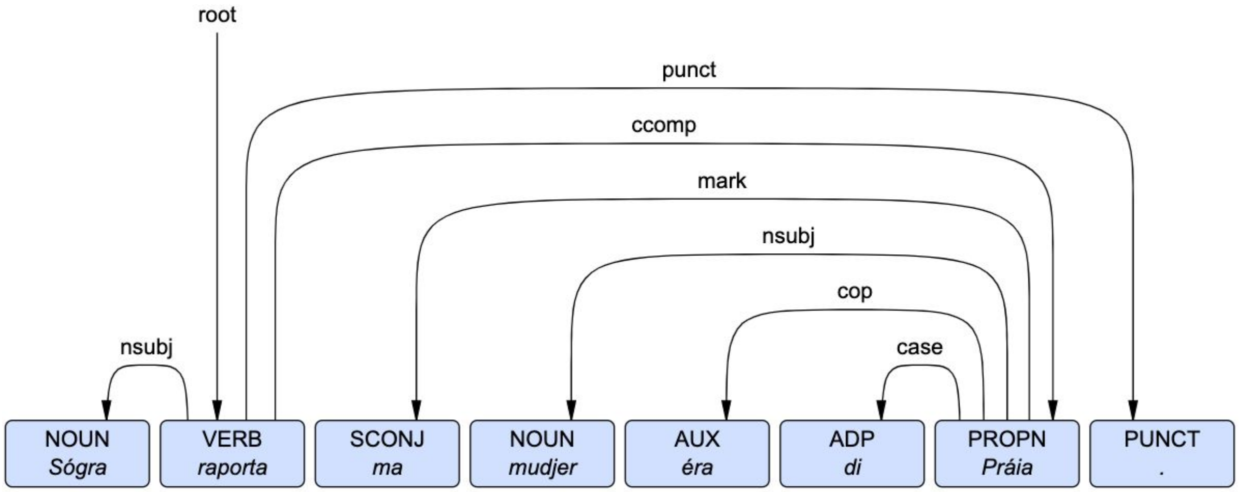
21. **Flat Multiword Expression (flat)** - Used to describe headless semi-fixed multiword expressions such as names and dates.



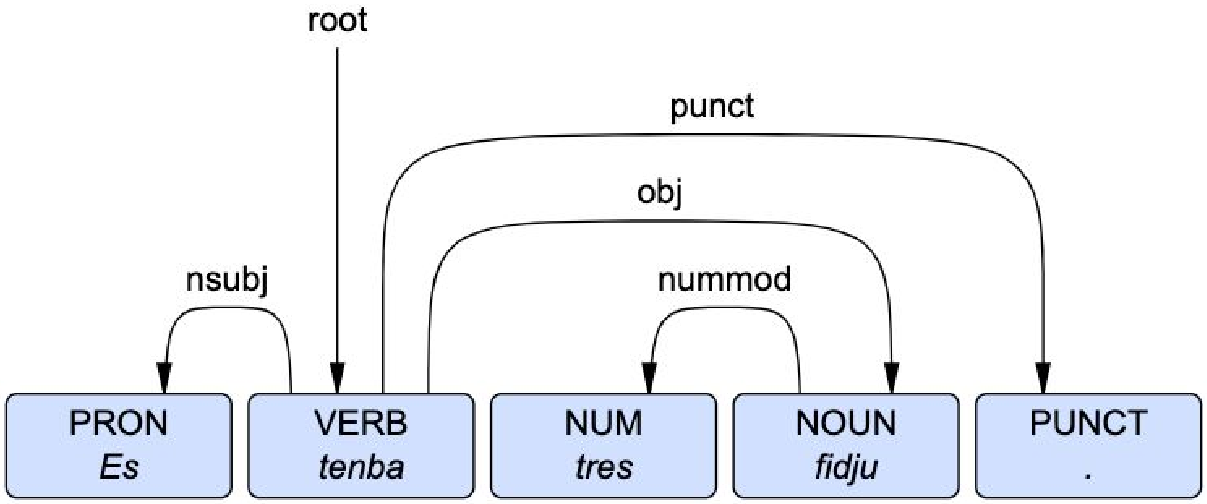
22. **Goes With (goeswith)** - This relation links two or more words that were supposed to have been written as one word. The head is always the first "word" in the sentence.
23. **Indirect Object (iobj)** - Typically the recipient of ditransitive verbs. The indirect argument is a core argument of a verb that is neither the subject or object.



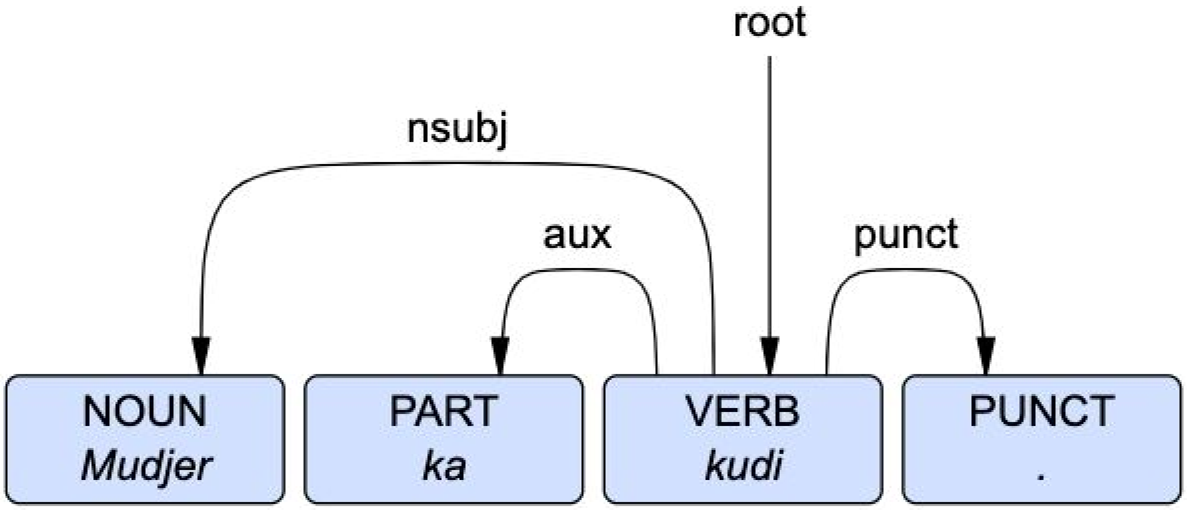
24. **List (list)** - This relation is used for lists or chains of comparable items. Often seen online for contact information or other chunks of related information.
25. **Marker (mark)** - A word which marks one clause as the subordinate of another.



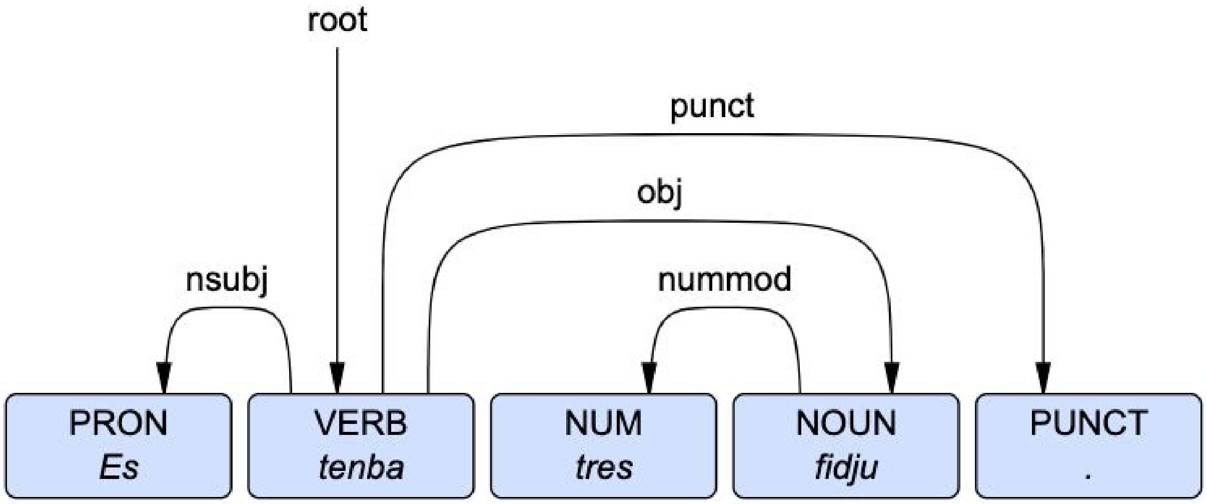
26. **Nominal Modifier (nmod) -** Used for nominal dependents of another nominal, and the function is usually to describe or convey a genitive attribute.



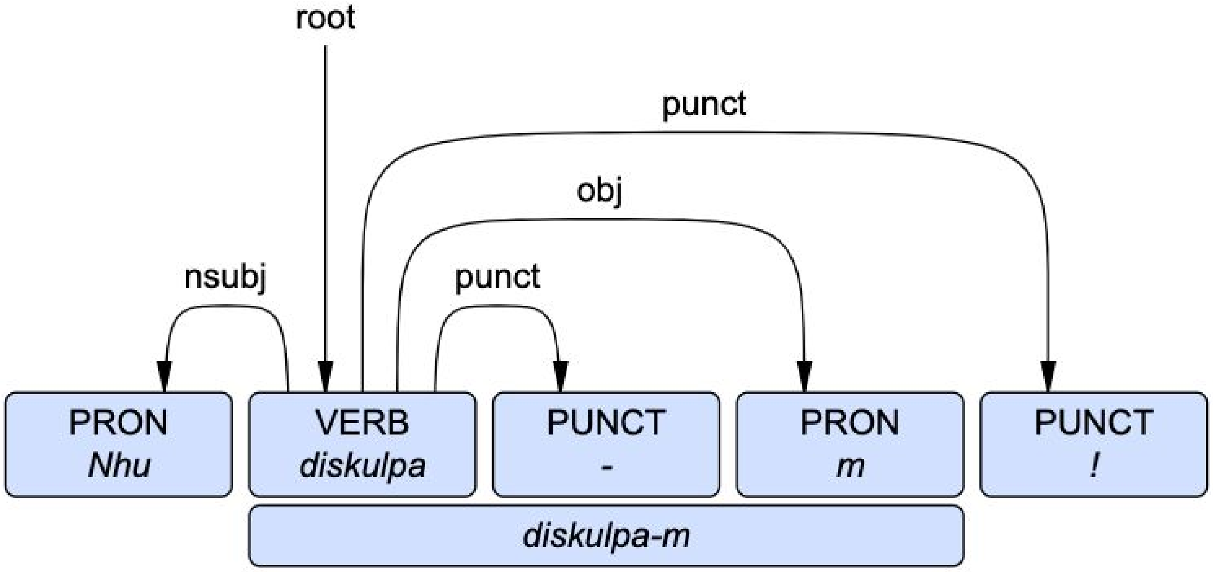
27. **Nominal Subject (nsubj) -** This relation is used to describe a nominal which is acting as the subject of a verb.



28. **Numerical Modifier (nummod) -** Used to connect a number, in digital or textual form, which describes the quantity of a nominal.

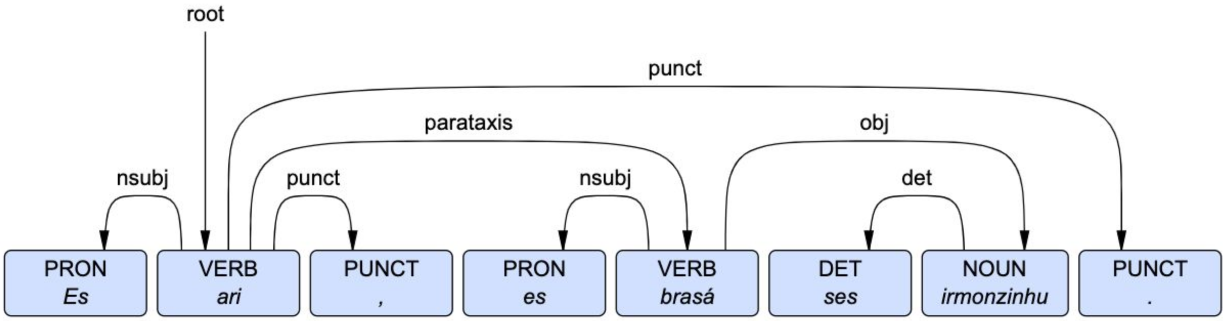


29. **Object (obj) -** The secondary core argument of a verb.

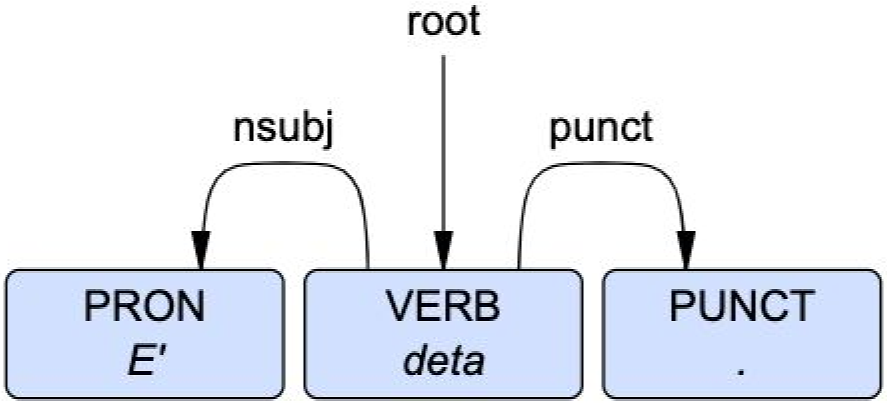


30. **Oblique Nominal (obl) -** Describes a nominal acting as a non-core argument of a verb, in certain situations, it is used to express case.
31. **Orphan (orphan) -** This relation is used when there is head ellipsis and simply promoting would lead to an unnatural sentence.

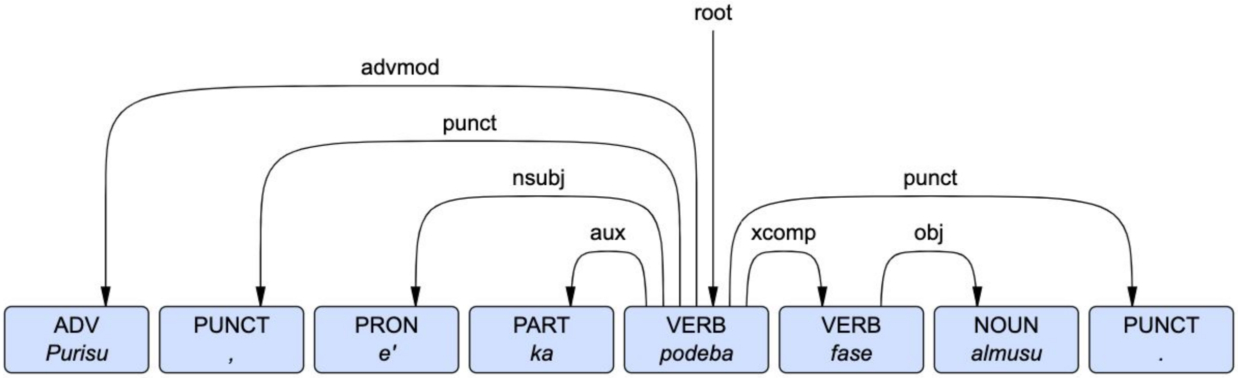
32. **Parataxis (parataxis) -** Used when independent clauses are found in the same sentence without coordination, the first clausal root is seen as the head.



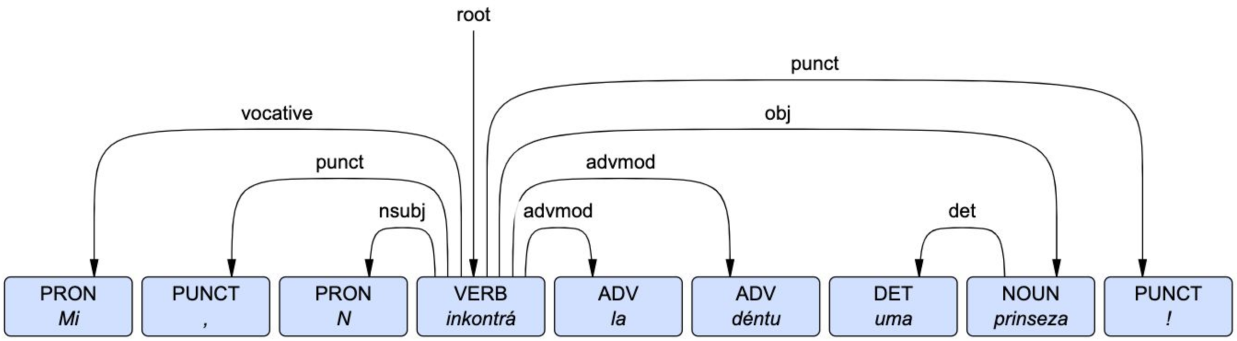
33. **Punctuation (punct) -** Used for any piece of punctuation in the sentence.



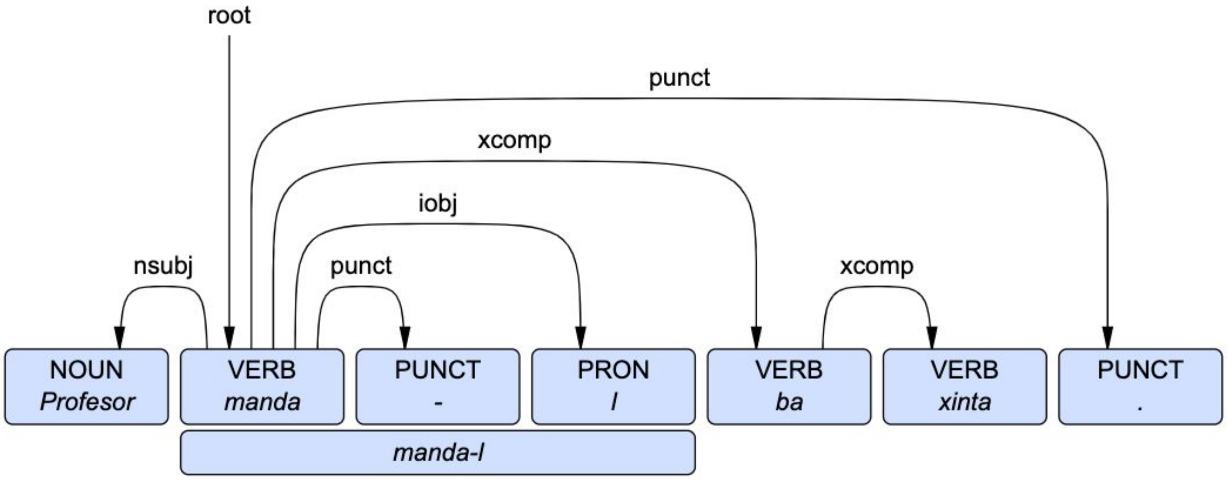
34. **Reparandum (reparandum) -** Used to denote disfluencies overridden in speech repair.
35. **Root (root) -** This relation is used to mark the main verb of the sentence, a pseudo node takes the role of the governor of the sentence and the main verb is its dependent.



36. **Vocative (vocative) -** Used to mark an entity being addressed directly in dialogue.



37. **Open Clausal Complement (xcomp) -** A clausal complement of a verb or an adjective which has no subject.



PARSING WITH UDPIPE

The end goal of manually annotating text in Cape Verdean Creole is to then be able to train a parser which is capable of taking in raw text in CVC and outputting the text tokenized, tagged for part-of-speech and also marked with the appropriate dependency relationships.

In order to train the parsing models for tokenizing, tagging and dependency parsing, I used the open-sourced UDPipe pipeline.

UDPipe is available as a binary capable of being used on Windows/Linux/OS X and using annotated data it allows us to train models without any language specific data. The pipeline only requires data that is annotated following the CoNLL-U format which adheres to the guidelines of the Universal Dependencies Project. Following the CoNLL-U format, words are separated by lines, and sentences are marked by an empty line. For each line for a given word, the information for each word is separated by a tab. There are ten fields for each word which correspond to:

1. **Word Index Number**
2. **Word Form Found in Text**
3. **Lemma**
4. **Universal Part of Speech Tag**
5. **Language Specific Part of Speech Tag (Underscore if Unavailable)**
6. **Morphological Features**
7. **Head of the Word**
8. **Universal Dependency Relation to the Head**
9. **Enhanced Dependency Graph**
10. **Miscellaneous**

Example:

```
1	fidju	fidju	NOUN	_	Number=Sing	3	obj	3:obj	_
```

From there, we are able to produce models such as a tokenizer, lemmatizer, part-of-speech tagger and of course a parser for dependency relations. The pipeline, by default, will train three models for a given dataset; one for tokens, one for tags and one for dependency relations. However it can be configured to only train one or two as well.

The UDPipe pipeline uses MorphoDiTa, an open-source tool for morphological analysis, for training the tagger, and *Parsito,* a transition-based parser which utilizes a neural-network classifier.

In order to train the model 400 of the original sentences were used while 42 sentences were used for development and another 42 for testing. Due to the size of the treebank, the training was only carried out over 10 epochs. The UDPipe default of 100 epochs would have been unnecessary with this amount of data.

A total number of 534 trees were analyzed, with 9300 words, 8659 tokens, and 39 dependency relations (with 4 being language specific).

The UDPipe pipeline evaluates the generated language model using both a labelled-attachment score (LAS) and an unlabelled (UAS) score. The two test slightly different aspects of the model. LAS corresponds to the accuracy of dependency labeling whereas the UAS score references the accuracy of the dependency structure, testing whether the correct head was assigned.

The current model for Cape Verdean Creole achieved a LAS score of 67.47% and a UAS score of 74.47%. The evaluation for the tagger resulted in a score of 82.98% for all tags. Overall, I am happy with the results of the model training, especially given the size of the dataset.

CONCLUSION AND FUTURE WORK

Moving forward, the first step would be to increase the size of the dataset inorder to train a more robust model. Due to the text coming from the same book, the dataset was skewed to the writing style of the author. The next steps of building onto this treebank would be finding more data from various sources.

However, creating one large model that can parse all of Cape Verdean Creole text is a very daunting task however given the previously mentioned high level of variability in spelling and grammatical conventions.

We even see different spellings used within the same chapters of the book by T.V da Silva. It can be almost guaranteed that when looking at texts written by authors of different islands and time periods, we will see even more variability.

There is also a noticeable difference between the Barlavento and Sortavento dialects that might require two separate datasets.

Prior to using text from various sources, which would be critical in order to build a more diverse dataset that is a true representation of the way the language is used across the archipelago and in the diaspora, a spelling paradigm would have to be agreed upon and text not conforming to that would have to be altered prior to annotation. Otherwise, we would have a dataset rich with different spellings of the same word, which is not optimal for training a model.

This treebank for Cape Verdean Creole is a useful resource for future linguistic study, but it also outlines the challenges that need to be faced in order to integrate the language into cross-linguistic study. Until the language is standardized and formally taught in schools, I believe that this challenge will persist due to a lack of a large uniform body of data.

In the end, I believe this work is a contribution to the future of Cape Verdean Creole in natural language processing as well as to the overall research of the language of the people of Cape Verde.

BIBLIOGRAPHY

I. Universal Dependencies - Part-of Speech Tags. https://universaldependencies.org/u/pos/index.html

II. Universal Dependencies - Dependency Relations. https://universaldependencies.org/u/dep/index.html

III. Universal Dependencies - Morphological Features.https://universaldependencies.org/u/feat/index.html

IV. Universal Dependencies - Format. https://universaldependencies.org/format.html

V. T.V da Silva, Na Boka Noti

VI. Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection 2016

VII. Ryan McDonald, Joakim Nivre, Yvonne Quirbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, Jungmee Lee. Universal Dependency Annotation for Multilingual Parsing 2013

VIII. Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slave Petrov. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies

IX. Daniel Zeman, David Mereček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, Jan Hajič. HamleDT: To Parse or Not to Parse? 2012

X. Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. Universal Stanford Dependencies: A cross-linguistic typology

XI. Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, Larraitz Uria. Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies

XII. Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, Valeria de Paiva. Universal Dependencies for Portuguese

XIII. Teresa Lynn, Jennifer Foster. Universal Dependencies for Irish

XIV. Elena Badmaeva, Francis Tyers, Koldo Gojenola, Gosse Bouma. Universal Dependencies for Buryat

XV. Noam Chomsky. Syntactic Structures 1957

XVI. Milan Straka, Jana Straková. UDPipe. http://ufal.mff.cuni.cz/udpipe

XVII. CoNLL-U Viewer. http://www.let.rug.nl/kleiweg/conllu/

XVIII. Slav Petrov, Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman. Syntactic annotations for the Google Books Ngram Corpus 2012

XIX. Manuel Da Luz Gonçalves. Cape Verdean Creole to English Dictionary