Project Proposal: Building a South African English corpus and Assessing Part of Speech tagging accuracy

Alec Badenhorst
bdnale004@myuct.ac.za
University of Cape Town
Rondebosch, South Africa

Muhammed Umar Khan
khnmuh038@myuct.ac.za
University of Cape Town
Rondebosch, South Africa

KEYWORDS

Natural Language Processing, Corpus, Part-of-speech, Named Entity Recognition

1 INTRODUCTION

Natural language processing (NLP) is a field of computer science that has been widely applied to various languages around the world with great success in many. English in particular has been the target of much language processing. English however comes in many variations (dialects if you will) from different countries. Some of the most widely known English dialects are American, British, and Australian English. NLP has been applies to all three of these with very great success. South Africa has developed its own English over the years, taking influence from many of the local Bantu languages, and Afrikaans. Despite this, little research with respect to South African English has been done up to this point. In particular we wish to tackle the problem of building a Corpus of South African English, and verify the efficacy of existing part-of-speech taggers on South African English, and if they are as good with South African English as with other variants, tackle the problem of named entity recognition.

Natural language processing (NLP) is a field of computer science concerned with the processing of natural (i.e. spoken) language by computers. NLP has various applications in the real world including machine translation, and understanding of language when spoken to (i.e. Siri on an iPhone). NLP generally uses a corpus to train, such as the TreeBank corpus [14], which was built using the TreeTagger [20].

A corpus is a collection of texts which are used to study a language and other linguistic findings.South African English which contains multiple sub varieties such as Black South African English, Indian South African English, Afrikaans South African English e.t.c has no large scale electronic corpus available [2].

The need for a South African English corpus comes in the form that the International corpus of English or ICE requires that there be a South African English corpus as it is one of the components of the ICE [5].The development of a South African corpus is further emphasized by the fact that an authority of SAE the Dictionary of South African English (DSAE) requires that a SAE corpus be available because their current approach of manually collecting data to compile SAE dictionaries seems to limit the amount of data that can be gathered and also hinders to what depth the analysis of data can take place[2]. An SAE corpus needed by the English Language and Linguistics Department of Rhodes University [2].

There has been work done in developing corpora for SAE but these corpora focused on sub-varieties of SAE such as Indian SAE or Black SAE [2]. The sample sizes for these corpora were relatively small as the collection for data for these corpora involved the use of audio recordings [9, 5] making these corpora not useful for large scale projects.

Part-of-speech tagging assigns a part of speech (verb, noun, adjective, etc.) to words in a sentence. This is done through various techniques such as using rule based taggers, Hiddem Markov Models or maximum entropy models. These each have their own advantages and disadvantages, and deal with certain cases differently, producing potentially different accuracies. Part-of-speech tagging is one of the earlier steps in language processing, and is one of the more important steps, as knowing the part of speech a word represents is important to understanding the meaning of the entire sentence.

A sub field of part-of-speech tagging is named entity recognition. Names entity recognition is the determining of so-called named entities in a sentence. The most straight forward example of a named entity is a person's name. The person is a named entity referred to by their name in a sentence. The word 'it' in a sentence could also be a named entity, as it is referring to something used in the conversation without reusing the same word again. Named entity recognition has some struggles when it comes to South African English. Some adjectives have been used as names in some cultures in South Africa. Some of these names include 'Pretty,' 'Precious,' and 'Grace'. In the middle of a sentence, determining

if something is a name could be a simple as looking out for capitalization, but if it is at the start of a sentence, it might become more difficult for taggers to recognize.

Taking the points raised above into account, it the need for a corpus of South African English is necessary to further NLP research with respect to South African English. A corpus of SAE will help further research in other areas such as part-of-speech tagging and named entity recognition as well. The need for accuracy in part-of-speech tagging and named entity recognition are also presented and make clear why it is important.

2 PROBLEM STATEMENT AND RESEARCH QUESTIONS

Part of speech tagging is largely considered a solved problem when considering English. People consider it solved since most modern taggers reach similar accuracies of 96% [17]. Although this is true for most Englishes around the world, it is unclear whether this holds true for South African English. A sub area of part-of-speech tagging is named entity recognition. Named entity recognition doesn't quite reach the accuracies of general part of speech tagging, but suffers significantly when it comes to South African English. [17] reached F-scores of up to 90%. When compared to existing research in the South African context, our 75% falls short [3, 12]. A probable reason for the lack of accuracy in the South African context could be ascribed to the lack of official gazetteers, as well as mixing of several languages.

As stated in the previous section previous work has been done to develop a SAE corpus, but these projects mostly focused on sub varieties of SAE which can differ quite a bit between one another and still pose the challenge of how these corpora for the sub varieties can be merged to help form a much bigger and comprehensive corpus for SAE. At the end of the day we are still faced with the major problem that currently there exists no large scale electronic corpus for SAE [2].

There are three goals for this project.

(1) To develop a comprehensive corpus of South African English that may be used computationally.
(2) Perform an analysis of various part of speech taggers performance when presented with South African English.
(3) To develop an improved part of speech tagger or named entity recognizer for South African English.

2.1 Development of a South African English corpus

The main aim of this section would be to be able to develop a comprehensive and varied corpus for South African English, we would like to incorporate the various sub varieties of South African English fairly and would like to build such as system that future work and research can be easily conducted with it.We also aim to build such as system that will incorporate different genres of text each of which can enable further research and expansion.

So are aims for this project would for be.

(1) Build such a corpus for SAE that it could easily be expanded upon in the future and be used for research by the end of this project.
(2) Applying methods used in the construction of other language specific corpora to develop a SAE corpus .
(3) Development of an appropriate metadata scheme for the corpus, this will aid in adding linguistic meaning to the corpus and make it easier to navigate the corpus and extract information from the corpus.

2.2 Part of speech tagging of South African English

We wish to tackle the problem of the efficacy of existing part-of-speech taggers on South African English. This project aims to answer four questions. Can existing part of speech taggers tag South Afrian English with a similar efficacy to other English variants (within a reasonable margin of error)? Which part of speech tagger is the most effective on South African English? Can we improve one of the taggers specifically for South African English? How does the domain of application affect the efficacy of part of speech taggers?

The main goal of the project would be to analyse the effectiveness of different part of speech taggers on South African English and compare their performance when applied to South African English to their performance when applied to other varieties of English. We know the efficacy of part-of-speech taggers with respect to English is very good, as shown by [6, 8], however, there is no definitive study that tackles the efficacy of part of speech tagging with respect to South African English specifically. Schlünz, Dlamini, and Kruger [18] made a reference to the efficacy of part-of-speech tagging with respect to South African English, but the language used in their data set was of a legal nature, and thus might not have included very many South African English specific words.

In doing these experiments, we wish to determine which

techniques or tools end up being the best. If it is deemed appropriate, a specific tagger might be modified and improved upon to increase accuracy for South African English.

If the taggers are deemed to be as good as it gets (reaching up to 96% accuracy consistently [8]) the focus of this project may be redirected to named entity recognition, a sub field of part-of-speech tagging. Difficulties with named entity recognition and South African English has been noted [18], and is definitely an area of research that deserves more attention.

3 RELATED WORKS

In this section we will be focusing on the related works that have been carried out with part of speech tagging accuracy and the development of a SAE corpus or methods used in construction other language specific corpora that can be implemented for developing a SAE corpus.

3.1 SAE and other language specific corpora development procedure

To develop corpora the following process was generally followed by authors:

(1) Gathering data for the corpus
(2) Storage of acquired corpus data
(3) Filtering and cleaning acquired data
(4) Annotation of the corpus

3.1.1 **Gathering corpus data.** . Dwyer [2] specifically worked on developing a corpus for South African English and the methods used by [2] was the use of web crawlers, RSS feeds and the collection of dynamic data after short periods of time.

[1] and [13] which constructed large corpora used web crawls as well but also highlighted the possible use of methods such as search engine queries and search engine hit counts.

3.1.2 **Storage of acquired corpus.** Many scholars generally tended not to state where they stored the corpus [2].Dwyer [2] used NoSQL databses such as MongoDB to store all data retrieved for the corpus and scholars such as [11] chose to store data locally in flat text files as they believed this method was much more desirable than traditional relational databases.

3.1.3 **Filtering and cleaning acquired data.** Once corpus data was collected it required filtering and cleaning, procedures such as boilerplate removal had to be carried out by [2, 13, 1] and either developed or used specialized tools for this process such as BTE, NCleaner etc.During the cleaning process scholars also performed the removal of near duplicates or exact duplicates of data. Exact duplicates were simpler to handle as with the use of hashing and fingerprints to track HTML documents [2].Near duplicates was a harder task to accomplish but the general approach seen with many scholars such as [2], [13] and [1] was the use of of n-grams to track the similarity between documents and also implement something known as the "shingling" algorithm to deal with near duplicates [2, 1, 13].

3.1.4 **Annotation of the corpus.** Corpus annotation had to be carried out to make the corpus actually useful for linguistic research. Open source tools such as IMS Corpus workbench were used used by [13] to index the corpus that they had built.Dwyer [2] used NLTK and their default Penn tagger to perform POS tagging of the data.Most scholar focused on forms of POS tagging when it came to annotation of a corpus.Lijffijt.i [10] separated the corpus data to different genres or categories by monitoring aspects such as average word length and sentence length in different data sources.

3.2 Part of speech tagging and South African English

Much research has been put into the field of part of speech tagging and there are several methods by which part of speech tagging takes place. We wish to determine if any existing taggers are able to tag South African English sufficiently before deciding on the next step. Part-of-speech tagging may not be quite as effective for South African English as for other variants of English due to the presence of many loan words, be they from Afrikaans, any of the Buntu languages, or other languages present in Southern Africa. Schlünz, Dlamini, and Kruger [18] made a reference to part-of-speech tagging for South African English using the HunPOS [4], but their domain is unknown. Assuming that their numbers reflect the expected accuracy, the focus would be shifted to named entity recognition.

Louis, De Waal, and Venter [12] have applied named entity recognition to South African texts. They used some rule based systems to determine whether something is a proper noun or not such as checking for "ś" or capitalization in the middle of sentences. Louis, De Waal, and Venter [12]'s research excluded common names that are also adjectives such as "Precious." or "Gift" as their tagger could not deal with these. They achieved F scores (measure of accuracy) between 0.42 and 0.67 without and with a gazetteer (a catalogue of names) respectively. Official freely available South African gazetteers don't exist, so [12] took names from US census data combined with South African names taken from employee lists. Louis, De Waal, and Venter [12]'s approach was to use a dynamic bayesian network, which represent

the probabilistic relationship between a set of variables at some point in time [15].

Eiselen [3] has also applied named entity recognition on South African English texts, although their focus was rather on South African languages as a whole rather than South African English. They focused their research on government domain, and achieved F-scores of roughly 0.75 for most South African language, including our language of interest, English. Eiselen [3] followed general annotation principles in guiding their decision as to whether something was a named entity or not. There are four principles;

(1) The token must be a unique identifier.
(2) The token must be a proper name, most likely written with capital letters.
(3) The name of the entity must be assigned through some official process such as a birth certificate, official registration, or an assignment through a law or governmental agency.
(4) In figures of speech where the exact type is unclear, the most prototypical interpretation is assigned.

4 PROCEDURES AND METHODS

In this section we discuss the Methods and design aspects of our respected aims.

4.1 South African English corpus development

To carry out the corpus development process we will break it up into the 4 major parts as mentioned in section 3.1.

4.1.1 **Gathering corpus data.** Various methods were considered when it came to deciding the methods that were going to be used to gather data for the corpus, these included search engine queries, search engine hit counts, web crawlers and RSS feeds.

We decided to use web crawlers for one of the methods for data collection, crawlers were used cause they are efficient at gathering data and there are many open source crawlers available to use, web crawlers also don't have limitations that are present with methods such as search engine queries, where the number of queries per day is restricted by search engines. The Heritrix crawler will be used as it prioritizes large amounts of data and is designed to fit many use cases. Another web crawling tool that will be used will the Scrapy crawler which a python based crawler, Scrapy also allows for the easy creation of filters for the cleaning of data as well as data sanitization. To ensure we are dealing with SAE we will be crawling South African websites such as government websites, South African newspapers, Social media pages, and south African versions of popular sites such as Google or Yahoo.

Another method we want to explore is being able to retrieve data that changes dynamically. We believe this is specifically important when it comes to social media sites or any place where comments or chat rooms are, We believe gathering such data will be helpful to retrieve words used by the general audience and the general structure of the public sentences.Use of open source software such as the Crawljax will be used to extract data from these comment sections.

To ensure we are dealing SAE text we will be crawling web pages such as public social media pages such public Facebook pages and be crawling users comments that will represent the English used by the general audience, when crawling public social media web pages we will also look at pages that are targeted towards a particular region or group in South Africa an example might be a a web page targeted towards people of Cape town in which we expect Afrikaans English and colored South African English to be prevalent.

4.1.2 **Storage of Acquired Data.** We also need to come up with an efficient way to gather the data we acquired from crawling, crawljax etc. Corpus data storage will be done with the use of flat text files.

We will make the use of flat text files as done by [11] which had a index file to track data.This approach was chosen over RDMBS as it scales well and is not so resource intensive as RDMBS systems.

4.1.3 ***Data cleaning and filtering***. After gathering and storing data, the data is still far from ideal to be used for the development of a corpus.There exists problems such as boilerplate and data duplication that need to be taken care of.

Our approach for the removal of boiler plate would be to use open source and free programs such as Boilerpipe(used by [2]) and BTE (used by [1, 7, 13]) these programs are automatic cleaning tools that are used for the removal of boilerplate, if weakness is seen with these approaches a custom Python script will be developed to aid in the removal of boilerplate.

The problem of data duplicates will have to be divided into two subsections.The first being that of exact duplicates in which we saw the most common approach was the use of some sort of hashing or fingerprint algorithms as used by [2] and seemed to get rid of exact duplicates rather well with the use of MD5 and SHA hashing algorithms, so the same types of algorithms will be implemented t.However, dealing with near duplicates is a much more complex issue and will

be dealt with by using methods such as the "shingling" algorithm used by [13] which involves the use of n-grams to get rid of near duplicates.So will be implemented at a sentence level for each document and generally if the grams match for than 50 percent then the documents will be discarded the reason such a high percentage is chosen is because of an example stated by [2] in which if different newspaper articles publish the same story, so in that case obviously there will similarities between the newspaper articles.

4.1.4 ***Corpus Annotation***. This step will be required to solve our aim to actually be able to use the corpus for the purposes of linguistic research. We believe initially it is important to divide the data gathered into their various sub sections or genres to make it much more easier to extract specific type of data from the corpus e.g extracting data about only South African newspapers.

We would like to implement a similar technique to [10] where genres can be separated by monitoring the features such average word length, average sentence length, frequency of nouns and pronouns e.t.c.We aim to use these features to separate the data collected into their respective categories or genre.

Thereafter we would also like to split the text according to their respective province this includes for example a local newspaper from Durban which is expected to comprise of Indian South African English or Black South African English which in this scenario would be spoken among the Zulu population. We believe this is necessary as South African English has many sub varieties which are used in various regions throughout the country, so we feel that separating text according to their province will help with identifying the sub varieties of SAE.

We also need to test if the method used by [10] holds true for South African English this will be done by using a small sample size first in which we will use a few examples from various categories such as political meetings, newspaper, university lectures and public conversations between people.

At this stage we also need to take into account implementing sub-genres as for example a local public announcement or a local counselors speech might be very different from the speech used when a politician is giving a speech intended for the entire nation.The same can be said about small regional newspapers or newspapers targeting a specific audience, the text contained in these newspapers is expected to be different from large newspapers intended to be read throughout the country.

4.2 Testing strategy for SAE corpus during development life cycle

An iterative approach will be taken with respect to the SAE corpus development.The stages involved during our corpus development will each have their own set of tests for example when gathering the corpus data we will need to evaluate the efficiency of the crawls produced by both Heritrix and Scrapy, if by chance both the crawlers do not work well, then we will look for alternative web crawlers that fit our requirements. The same approach will be used for the other stages of corpus development testing the programs or methods in each case to make sure the development process is on track, so, for example during the annotation section initially we can just focus on one genre or category of data like written text and spoken transcribed recordings of political conferences or meetings, this will help us to make changes to our design easily rather than testing everything right at the end which will then make it very difficult to make changes to the overall design.

Our strategy is to first create a smaller corpus in the first few weeks and test it before diving into the developing the full SAE corpus, this could include maybe just gathering data from a few government websites and see how we can apply the steps required for corpus development to that data. This strategy will also enable to start testing the efficiency of POS taggers on SAE early on during the corpus development and the further testing of the POS taggers can be done as the development of the corpus goes on and more text is added.

4.3 Part-of-speech tagger efficacy on South African English

We wish to determine which of the taggers are the most effictive at tagging South African English, and if they excel at different domains. To answer our questions of efficacy and domain, the experiment will be conducted into phases. Each phase will have an iteration of each of the different taggers we will be using. Each phase will involve the testing of the taggers against a specific domain of South African English, e.g. government, news, novels, etc.

To determine the efficacy of existing taggers on South African English, we will use several South African English pieces of text from different domains, and use several different taggers on these texts. Various taggers for parts-of-speech exist online, many of which are free and open source. These taggers will each run on the same texts, and their accuracy will be measured. The accuracy of the taggers will be measured using an F_1 score, a statistical measure of the accuracy of the tests. To determine the F1 score, the taggers will be

run against an increasingly larger corpus provided by Umar. Starting small, we can manually check the efficacy of the taggers which allows us to calculate the F1 score.

State of the art taggers reach 96% accuracy [8] when tagging English texts from American and British English. If the taggers tested here all reach similar results, the focus of the project will be shifted onto named entity recognition using the tagger which performed the best. If all the taggers falls short, the focus of the project will be to improve upon the its tagging capabilities and build a tagger specifically made for South African English. According to Schlünz, Dlamini, and Kruger [18], tagging South African English gives results comparable to other English variants. They applied the freely available HunPOS tagger [4], but the domain of application is unclear. Given this result, it is expected that tagging for South African English will already be as good as it gets and the focus of the project will be shifted to improving upon named entity recognition for South African English.

In either case, an open source tagger will be used and improved upon during this project.

The final improved tagger (be it for named entities or parts of speech) will be developed in python using NLTK [16]. Project [16] is a python library and a platform on which to build NLP software. Project [16] includes the Tree tagger [20], HunPos [4], as well as the Standford tagger [19], making it an ideal platform to use for performing the necessary tests as well as improving upon an existing tagger or named entity recognizer.

The performance of the final product will be compared to state of the art taggers with known accuracies, such as Ratinov and Roth [17]'s, or Kupiec [8]'s. Since we know what kind of accuracies are possible for English, we have a goal to strive towards. The minimum we wish to gain out of this is an improvement upon the existing taggers for South African English such as Louis, De Waal, and Venter [12]'s and Eiselen [3]'s performances. Louis, De Waal, and Venter [12]'s named entity recognizer had obvious shortcomings such as ignoring adjectives, such as "Precious," "Gift," and "Grace," that are used as names.

Several taggers exist which will be made use of during this experiment. These tools include

- TreeTagger [20]
- HunPOS [4]
- Stanford POS tagger [19]

Some of these are not open source, but are still free to use. They were chosesn specifically due to their maturity and efficacy on American and British English.

Modification will be done on an open source one if possible, otherwise a tagger may be built from the ground up. Project [16] allows for one to build a part-of-speech tagger (and by extension a named entity recognizer) from scratch with its library, as one could incorporate it into a machine learning algorithm, and train it on existing corpora.

5 ETHICAL, PROFESSIONAL AND LEGAL ISSUES

For this project we do not believe any ethical or legal issues should arise, for the corpus development process all tools that will be used will be free and open source, a potential issue could be using copyrighted data for the corpus but all sources where data will be retrieved from will be cited properly and if copyright text is used then that section of the corpus will not be published. The final corpus produced will be open source for the most part as we aim to use copyleft text for the construction of the corpus.

When it comes to assessing part of speech tagging for South African English the only issue to keep in mind is not to violate the license agreement of any of the taggers that will be used to perform experiments on text.

All code that will be written for the project will be released under the creative commons license, this hold true for the publishing of the final project paper as well.Modified code on the other hand will retain its original license agreement.

6 ANTICIPATED OUTCOMES

In this section we cover what we expect to see by the end of the project, the impact the project will make and what what key factors we will use to judge the success of the project.

6.1 System

For the development of the SAE corpus the anticipated outcome is to overcome multiple challenges faced when developing the corpus for SAE, so that the development of a SAE corpus may be achieved. As stated earlier the process will be broken up into 4 parts: gathering data, data storage, cleaning ad filtering data and corpus Annotation.

Key features of the corpus will include a corpus separated by genre or category and the ability to locate keywords or specific types of articles.The corpus should also be able to be used by linguists with ease for research. Software used for the corpus development will be free and open source, with, if need be, some programs will modified and written to aid

with development process.

The biggest design challenge will be deciding the size and scope of the corpus that will be built.Deciding what data should be kept during the filtering and cleaning step of corpus development is also a concern as discarding some articles as near duplicates might lead to loss of important data.

We will have a definitive answer as to whether existing part-of-speech tagger are as effective on South African English as it is on other variants, and have built an improve tagger for either part-of-speech or named entities based on those results.

A design challenge in the improvement of a tagger will be in the form of adapting existing code, or even writing a new tagger from the ground up. If a tagger is to be built from the ground up, time constraints (due to training times) may limit the accuracy of the tagger.

6.2 Expected Impact

We believe that if we can develop a comprehensive SAE corpus, then the impact the project will have is that it will aid in linguistic research with respect to South African English. The corpus could also be further improved and be made a component of the ICE which is still under development and requires a SAE corpus.

Additionally, an effective tagger for South African English will help further natural language processing research in the domain of South African English. This could be applied to various important domains, notably that of government, but also be used to build improved machine translating tools, as well as text-to-speech synthesis.

6.3 Key success factors

- To develop a comprehensive corpus for South African English which can be used for linguistic research and be further expanded upon in the future, this will be achieved if various categories of text is contained within the corpus covering multiple fields in which different forms of SAE is used.
- Final testing of the corpus yields expected results which is to have a SAE corpus which encompasses many of the sub varieties of SAE, each represented fairly.This will ensure that the requirements set out earlier have been fulfilled.
- Definitive answer to the efficiency of taggers on South African English.
- An improved tagger for South African English and named entities.

7 PROJECT PLAN

We provide detailed information in this section on how we plan on conducting this project.

7.1 Required Resources

The data used for the development of the corpus will be taken from the web and transcribed audio recordings, this will include data from South African government websites, South African newspapers, South African social media pages and so on.

The only hardware resources that will be needed would be access to personal computers and an internet connection when it comes for software resources multiple programs will be implemented with all of them being free and open source.

For corpus the software that will be used it given below:

- Scrapy and Hertrix web crawlers
- Disqus API and Crawljax for dynamic data
- MongoDB for data storage
- BTE and Boilerpipe for cleaning data
- POS taggers such as NLTKs Pennbank tagger or the Tree tagger
- IDE for programs that will be written to aid with corpus development
- HunPOS
- Stanford POS tagger

7.2 Risks

We were able to identify various risks that may hinder the completion of this project, we have provided mitigation strategies for these Risks as well how to monitor to prevent these risks to emerge. We also provided management strategies to control the damage caused if a risk actually occurs.A detailed breakdown of this is given in the risk matrix in Appendix A.

7.3 Timeline and Milestones

The project will commence from now till the 19th of October 2020 with the submission of the project poster.A Gannt Chart has been included in Appendix B to give a detailed breakdown of the project timeline.

A list of the major project milestones is given below.

- Submission of project proposal and completing the proposal presentation.
- Feasibility demo presentations
- Implementation of a mini corpus
- Tests done of mini corpus and each stage of corpus development

- Efficacy of taggers on South African English tested.
- Development of improved tagger or named entity regocnition software.
- Submission of project paper outline.
- Submission of project paper draft.
- Submission of final paper.
- Submission of project code.
- Carrying out the final project demo.
- Completion and submission of the project web page.
- Completion and submission of the project poster.

7.4 Deliverables

A list of the project deliverables is given below:

- A comprehensible and usable corpus for South African English
- Am analysis of the efficacy of POS taggers on South African English
- Project proposal
- Project proposal presentation
- Feasibility demo
- Project paper draft
- Final project paper
- Project code
- Final project demo
- Project Web page
- Project poster

7.5 Work allocation

Each of the project members will be assigned with their own major tasks. M.Umar Khan will be responsible for the development and testing of a South African corpus and Alec Badenhorst will be responsible for the analysis of part of speech taggers on South African English, and the improvement upon part of speech taggers or named entity recognition software based on the results of the tagger efficacy. The final report will be worked on by both members and Alec Badenhorst will also perform an analysis of the accuracy of part of speech taggers on the corpus developed by M.Umar Khan if time allows.

REFERENCES

[1] Bernardini Baroni Marco. "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation". In: *Language resources and evaluation* (2009), pp. 209–226.

[2] Gareth Terence Bryant Dwyer. "Towards Automated Creation and Management of a South African English Web Corpus". In: *arXiv preprint arXiv:1104.2086* (2014).

[3] Roald Eiselen. "Government Domain Named Entity Recognition for South African Languages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 3344–3348.

[4] HunPos. *HunPos*. URL: https://code.google.com/archive/p/hunpos/ (visited on 05/10/2020).

[5] Chris Jeffery. "On compiling a corpus of South African English," in: *arXiv preprint arXiv:1104.2086* (2003).

[6] Tao Jianchao. "An English Part of Speech Tagging Method Based on Maximum Entropy". eng. In: *2015 International Conference on Intelligent Transportation, Big Data and Smart City*. IEEE, 2015, pp. 76–80. ISBN: 9781509004645.

[7] Adam Kilgarriff and Marco Baroni. "Large linguistically-processed Web corpora for multiple languages". In: (2006).

[8] Julian Kupiec. "Robust part-of-speech tagging using a hidden Markov model". In: *Computer speech & language* 6.3 (1992), pp. 225–242.

[9] Vivian de Klerk Leela Pienaar. "Towards a Corpus of South African English: Corralling the Subvarieties." In: *Thirteenth International Conference of the African Association for Lexicography* (2018).

[10] Jefrey Lijffijt.i. "A simple model for recognizing core genres in the BNC". In: (2017).

[11] Qiuling Wang Liujun Zhao Weizheng Kong1 and Lihua Song. "Construction of power industry corpus based on data mining and machine learning intelligent algorithm". In: (2019).

[12] Anita Louis, Alta De Waal, and Cobus Venter. "Named entity recognition in a South African context". In: *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*. 2006, pp. 170–179.

[13] M.Baroni and M.Ueyama. "Building general- and special-purpose corpora by Web crawling." In: (2006).

[14] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank". In: (1993).

[15] Richard E Neapolitan et al. *Learning bayesian networks*. Vol. 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[16] NLTK Project. *NLTK*. URL: https://www.nltk.org/ (visited on 05/10/2020).

[17] Lev Ratinov and Dan Roth. "Design challenges and misconceptions in named entity recognition". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. 2009, pp. 147–155.

[18] Georg I Schlünz, Nkosikhona Dlamini, and Rynhardt P Kruger. “Part-of-speech tagging and chunking in text-to-speech synthesis for South African languages”. In: (2016).

[19] Stanford. *Standford Log-linear Part-of-speech Tagger*. URL: https://nlp.stanford.edu/software/tagger.shtml (visited on 05/11/2020).

[20] TreeTagger. *TreTagger*. URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (visited on 05/11/2020).

Appendices

A RISK MATRIX

Risk	Probability	Impact	Mitigation	Monitoring	Management
Unable to complete deliverables by final date	1	10	Use proper planning for deliverables and avoid working on things not required for the final deliverable.	Weekly meetings to make sure each member is on track with their requirements	Consult with supervisor and find a way to reduce the final scope of the project
Scope of the project being unreasonable (too big or too small)	4	1	Consult supervisor for a proper scope	Monitor the amount of work done by each member regularly to determine if the scope should be changed	Increase or decrease the scope of the project to an appropriate amount
Improving upon existing part-of-speech/named entity recognition algorithms prove too difficult	5	8	Spend time fully understanding the way an algorithm works	Meeting with supervisor ensuring that we understand the code	Consult with supervisor for ways forward.
Development of a corpus takes too long to be used for part of speech tagging or named entity recognition.	4	6	Building the corpus small texts at a time so that tagging training can be performed as the corpus grows.	Weekly meetings with teammate and supervisor.	Use a separate set of texts or corpora for part of speech tagging and named entity recognition training.
Tools specified for development of corpus show unexpected and poor results which lead to them being unusable.	3	7	Carry out tests early on during the project with selected tools so if any modifications need to be made they can be made in ample time.	Test tools used in each stage of the corpus development cycle.	Consult with supervisor to reduce scope of the project so more time can be focused on looking for alternatives
Teammate drops out of university	1	1	Maintain good mental health	Continual communication with each other. Avoid overworking.	Instead of relying on each others results, each project can function independently.

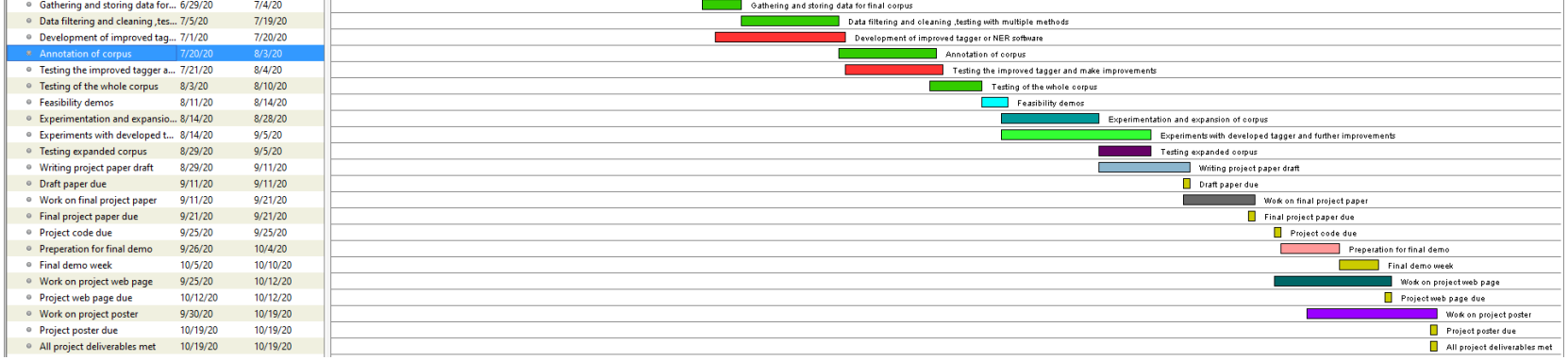

Figure 1: Gantt Chart

B GANTT CHART