NOVA IMS Information Management School

MDDM

Master's Degree Program in
Data-Driven Marketing

Automating Character Network Extraction for Portuguese Literature

Developing a pipeline for Literary Social Network Analysis

Tiago Gastão Gato Canário

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data-Driven Marketing

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AUTOMATING CHARACTER NETWORK EXTRACTION FOR PORTUGUESE LITERATURE

Developing a pipeline for Literary Social Network Analysis

by

Tiago Gastão Gato Canário

Master Thesis presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Digital Marketing and Analytics

Supervised by

Flávio Pinheiro, PhD, NOVA Information Management School
João L. M. Pereira, PhD, Universidade de Évora

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisboa, 15th July 2024

ACKNOWLEDGEMENTS

My deepest regards go to my supervisor, Professor Flávio Pinheiro, who guided me throughout this process, being ever present in times of need and without whom none of this would be possible.

To Professor João Pereira for his invaluable contributions and for challenging me to always take my work a step further.

Lastly, I am very grateful to have such a fantastic colleague as Catarina Ribeiro by my side during this process's trial and tribulations.

ABSTRACT

Social Network Analysis is becoming increasingly more relevant in the field of Literary studies. Achieving the automation of Character Network Extraction, a process that consists of mapping interactions between characters in a novel, so we are able to visualize the plot in a quantitative perspective, became a priority in the field. This enables analysis on vast corpora that would take months or years if performed manually. Automating this process requires the use of Natural Language Processing tools such as Named Entity Recognition and Part-of-speech tagging. However, these are not optimized for complex text present in novels and underperform due to the lack of manually annotated data for training, especially in less represented languages. For Portuguese, work was developed towards achieving an automated system but still, existing models rely on external information, which is usually unavailable, limiting their applicability to a large number of novels. Advancements in automated Character Network Extraction systems significantly enhance text mining capabilities in digital marketing by efficiently processing complex texts and mining user-generated content.

This work sets out to build the first fully automated Character Network Extractions system for the Portuguese. By adapting methods utilized in other works to Portuguese, we are able to overcome the suboptimal performance of Portuguese Natural Language Processing tools necessary for the automation process. Our results were tested on manually annotated chapters and compared to readily available State-of-the-Art tools that perform the same tasks, specifically an off-the-shelf Named Entity Recognition tool and the Large Language Model, ChatGPT. The resulting pipeline, which utilized Part-of-speech tagging and a combination of heuristics, achieved satisfying results with and average F1-Score of 94.1 in the task of identifying characters and solving for co-reference and 75.9 in interaction detection, an increase of 50.7 and 22,3 respectively on results achieved by the tools in comparison. Further steps to improve results are outlined, such as testing other solutions for interaction detection, and limitations on the size and scope of our testing samples are acknowledged. This pipeline was made publicly available to encourage development in this field for the Portuguese language.

KEYWORDS

Character Network Extraction; Social Network Analysis; Portuguese Language; Natural Language Processing; Text Mining

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Digital Humanities studies, with their quantitative approaches, are becoming ever more relevant in the field of Literary Analysis (Trovati, & Brady, 2014). Taking advantage of new processes of collecting, structuring, modelling and analyzing large volumes of raw data present in Literary novels opens opportunities to test previously unexplorable hypothesis and give new perspectives to old problems, having been used to disprove long-standing dogmas in the literary field (Elson et al., 2010).

Automating the Character Network Extraction process to conduct Social Network Analysis, has become a priority in the field due to its ability to tackle higher-level tasks and conduct analysis in exceptionally large corpora (Moretti, 2011). An example of the type of tasks that can be performed with this automation process is the study conducted on the sacred texts of the Chinese Buddhist canon that consists of fifty-five volumes and over two thousand additional texts and, as such, becomes impractical for any human conducted analysis (Lee & Wong, 2016). However, applying the required Natural Language Processing (NLP) techniques for automation to such ambiguous and diverse subjects as books has proven to be particularly challenging (Rocha et al. 2014). These tools are not optimized or intended to treat such complex text as the one present in novels and achieve suboptimal results in the necessary tasks (Labatut & Bost, 2019). The specificities of each language, such as grammatical rules or naming conventions, combined with the need of vast manually annotated corpuses used to train NLP systems to achieve satisfying results in literary text, left a big gap in these data acquisition systems for languages other than English (Bornet & Kaplan, 2017).

Previous works for the Portuguese language have tackled direct discourse in children's stories, utilizing heuristics to find instances of characters speaking (Mamede & Chaleira, 2004), developed systems that identify and extract information on characters present in a novel (Bick, 2023) or regard general entity extraction from non-fictional books (Rocha et al. 2014). The only existing Character Network Extraction System for Portuguese that we are aware of is dependent on external information sources, namely Wikipedia, to identify the characters present in a novel (Silva et al., 2023). As such, it is limited to the small universe of Portuguese novels that have that information available, and even those who have it concern only major

characters in the plot. A gap remains as there are still no systems that fully automates the Character Network Extraction process for Portuguese literature.

The advancements in automated Character Network Extraction systems have significant applications in digital marketing, particularly in the realm of text mining. The ability to process complex texts efficiently enhances the capability to mine user-generated content, social media interactions, and customer feedback, thereby providing a more comprehensive understanding of market trends and customer sentiments (Nassirtoussi et al., 2014). As these tools continue to evolve, they will play an increasingly vital role in shaping the future of data-driven and personalized marketing strategies (Saura, 2020).

This work sets out to answer the following research question: How can character network extraction methods be adapted to address the unique linguistic challenges in Portuguese Literary novels? The aim of this work is thus to build a pipeline that automates the Character Network extraction process that can be applied to a variety of novels. We achieve this by taking advantage of techniques employed in other languages, that improve results obtained using off-the-shelf and state-of-the-art tools (SOTA). Our resulting Character Network Extraction pipeline is made publicly available for further development and experimentation.

In this work, after conducting a thorough analysis of existing systems, both in English and in other less-represented languages such as French and German, a pipeline is built by adapting previously studied heuristics and other ruled-based approaches to the Portuguese language. To assess the pipeline's performance a corpus of chapters was manually tagged, and results were compared with those achieved by readily available SOTA tools that perform similar tasks.

2. LITERATURE REVIEW

This chapter aims to review the current state of Social Network Analysis (SNA) in literary studies. In subsection 2.1, we highlight the applicability of these methods and the work that has been conducted so far. In 2.2 we explore the theoretical challenges involved in developing an automated Character Network Extraction System both internationally and specifically within the Portuguese language. Lastly, in subsection 2.3, we examine the underlying principles of a standard Character Network Extraction system.

2.1. SOCIAL NETWORK ANALYSIS IN LITERATURE

Social Network Analysis and extraction, with its ability to uncover structural relationships and patterns within complex systems (Elsner, 2012), established itself as an interdisciplinary tool. It has attracted the attention of scientists, scholars, industries, and government agents that take advantage of its versatility, applying it in a variety of complex subjects (Choi & Kim, 2007). Web of Science database alone records 25903 articles using "social network analysis/data mining/text mining" in titles or abstracts between 2011 and 2020, covering topics from healthcare to energy and fuels (Fonseca et al, 2021).

The literary analysis field was quick to adapt and take advantage of this. Quantitative approaches have become increasingly more relevant in Digital Humanities studies (Trovati, & Brady, 2014). They utilize advanced computational techniques to analyze large volumes of text, uncovering patterns and insights that may not be visible through traditional qualitative methods alone (Elson et al., 2010). Consequently, more nuanced and data-driven insights can be drawn from literary works, thus, enhancing our understanding of literature (Labatut & Bost, 2019).

Distant Reading (Moretti, 2011), the quantitative evidence approach to literature, allows us to look at literary novels through a new holistic lens, where novels are seen as a whole, with a focus on a macro level and even allowing for the analysis of multiple books simultaneously (Lee & Wong, 2016). Not only do these tools allow for a level of abstraction to draw conclusions that were inaccessible before (Jayannavar et al., 2015), but they also enable us to perform, in short periods of time, investigations that used to take months or years (Moretti,

2011). Works in this field range from content to narrative structure, periods, genres and authorship clarification (Adanay & Sporleder, 2015).

The relevance of Social Network Analysis in the Literary field comes from its ability to structure character networks so they can be visualized by mapping out the interactions between characters in a novel to understand the complex relationships that drive the plot forward (Silva et al., 2023). In these models, characters are represented as nodes, while edges represent any type of relationship between them (Grayson et al., 2016).

Works developed in Literary Social Network Analysis can be organized into three main categories (Labatut & Bost, 2019), the first being Narrative Analysis. These focus on a small number of narratives at a time. This kind of study is performed to better understand a narrative through its plot, which is closely related to how the characters interact (Adanay & Sporleder, 2015; Sack, 2021). These analyses can be performed for a variety of purposes, to compare novels from the same author or period (Rieck et al., 2016) as well as test standing dogmas of the literary studies (Elson et al., 2010). There are also reports of these types of works being used in educational settings (Bolioli et al., 2013).

The second category of studies on Social Network Analysis in Literature are those of the Complex System model. These studies view the character networks present in Literary novels as intricate networks that mimic real-world social networks (Alberich et al., 2002; Bossaert & Meidert, 2013; Gessey-Jones et al., 2020) and try to find similarities between them, upholding the idea that understanding this networks, even if fictitious, can help us better understand real-world networks (Carron & Kenna, 2012; Trovati et al., 2014).

In the works from the two previously mentioned categories, analysis and conclusions can be drawn from a small sample of books or texts that can be manually annotated (Rieck et al., 2016).

The third category of work in the field falls into the Automation of the Character networks extraction process. Automating the process of extracting social networks from literature, enabling larger-scale projects and analysis, reproducibility of results, and turning an extremely time and resource-consuming process into an efficient and accessible one (Moretti, 2011) is what allows us to truly utilize Social Network Analysis as a tool to understand Literary novels.

It is important to note that these categories are not exclusive. One work can fit into two or more categories, as automation projects are usually built to conduct the previously mentioned analysis and test hypotheses (Elson et al., 2010).

2.2. Automation of the Character Network extraction and the Portuguese case

Achieving the automation of this Social Network Extraction is however a challenging task that poses many problems. The correct methodology for achieving faithful and relevant character networks has been the main source of debate within this field of study.

The challenge comes from natural language ambiguity (Rocha et al. 2014). The prose in literary novels is a form of unstructured text (Lee & Wong, 2016), with a complex nature both in form and meaning, proving to be so even for the human understanding (Zhai et al., 2013). All these situations, combined with other unique features that are highly specific to each novel, such as each author's individual writing style, create many issues in the automation of the extraction process (Labatut & Bost, 2019).

This complexity in the nature of literary text is even more accentuated when using Natural Language Processing tools required to automate this process. They are intended to apply and provide information in shorter and more direct text formats related to the real world, such as news articles and social media posts and are built around this principle. They take advantages of the traits present in this kind of text, such as context, to achieve optimal results (Bornet & Kaplan, 2017). Therefore, they are suboptimal when treating the text present in novels. These are written to portray self-contained worlds that can give you information only once or even do so in an implicit way and expect the reader to retain it, making it very difficult for NLP tools that depend on context and databases to correctly identify names.

Another trait of literary novels that hinders NLP tools' information retrieval capabilities stands from the intentional uniqueness in writing form that is intrinsic to each author, combined with the added variety of language mutations throughout the centuries (Dekker et al., 2019), wich significantly hurt the performance and capabilities of Natural Language Processing tools when handling a diverse corpus of literary novels.

Every factor previously mentioned contributes to the reduced performance of standard NLP tools such as Named Entity Recognition (NER) that are essential for automating the Character

Network Extraction Process. To solve this problem, NLP tools have been trained in large manually annotated corpora of literary novels, developed either as a collective effort from academic studies (Vala et al., 2015; Dekker et al., 2019) or achieved by crowdsourcing non-expert annotations (Callison-Burch & Dredze, 2010; Jha et al., 2010; Yuen et al., 2011) and made publicly available to further develop the field. However, the tools that originated from these efforts mostly concern the English language.

This situation contributed to a big disparity between the state-of-the-art tools in English and the ones for less-spoken languages (Silva et al., 2014). The less robust NLP systems combined with specificities in sentence structure, name and title rules that are unique to each language resulted in a higher prevalence of ruled-based approaches for the automation of character networks extraction in languages other than English (Bornet & Kaplan, 2017; Besnier, 2020). These procedures benefit from the ability to understand the logic behind the results and produce refinements to the pipeline that can be quickly adjusted in case of need (Krug et al., 2015).

Such is the case in work done in the Portuguese language. Due to the lack of robust NLP systems for the Portuguese language (Conrado et al., 2014), most works developed in the field are conducted using part-of-speech (POS) tagging and a combination of heuristics catered to the Portuguese names (Silva et al., 2023). The inability of NLP tools to tackle literary novels can be explained by the need for more initiatives for manually annotated corpora specialized in literary novels, a very time and resource-consuming activity. As previously mentioned, in some languages, this is addressed by taking advantage of crowdsourcing for mass annotations, a solution that is impractical for Portuguese due to the under-representation of workers who speak the language in this type of platforms. Even so, the interest in developing social network extraction systems remains, and some attempts have been made towards this goal.

Previous works in the Portuguese language have been developed for specific tasks within a Character Network Extraction System. Early work within the field laid the groundwork but was limited in resources and technology, tackling just direct discourse in children's stories (Mamede & Chaleira, 2004). A later work developed a more fleshed out system regarding general entity extraction from non-fictional books (Rocha et al. 2014). Two more recent works have shown progress in the area, the first being a system that identifies and extracts

information on characters present in a novel such as family relations and professions (Bick, 2023) and the second a full Character Network Extraction System that however relies on external annotations and information such as Wikipedia, utilizing a web crawler to retrieve information about characters names and relations (Silva et al., 2023). This limits the number of novels it can be utilized on since very few Portuguese books have the required information available in Wikipedia pages, and even the ones who do are at times incomplete, only regarding the most important characters.

As such, as far as we know, there is still need for an automatic character network extraction systems for Portuguese literature that can be applied to a large corpus of novels.

2.3. A STANDARD PIPELINE AND THEORETICAL CHALLENGES

As per the theory behind building an automatic social network extraction system, although pipelines for Character Network Extraction can differ depending on the intended goal of each study, besides general Character Network extraction for Social Network Analysis purposes, they can be focused on conversational networks and its topics (He et al., 2010) or even in extracting other kinds of information such as the type of relations established, e.g., friend or foe (Srivastava et al., 2016), there is a well-established three-step model for these data acquisition systems (Labatut & Bost, 2019) illustrated in Figure 1.

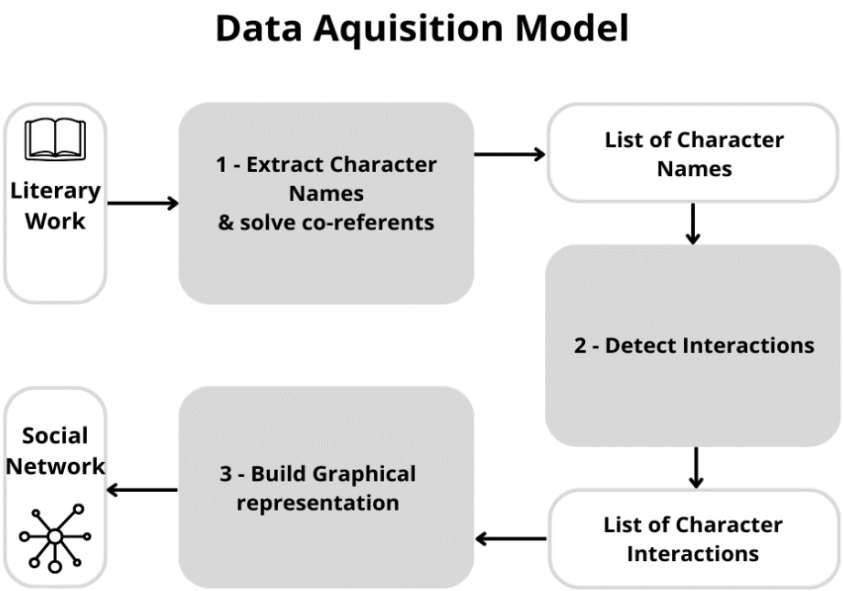


Figure 1 – Standard Data Acquisition Model for Character Network Extraction

The first step consists of identifying the characters present in a literary novel. During this phase, the final goal is to have a list of all the characters present throughout the text and identify every possible name used to mention the same character, a process designated by co-reference resolution (Vala et al., 2015).

Co-reference resolution has become so relevant in this field of studies that entire works are dedicated just to this specific problem (Vala et al., 2015; Bhattacharya & Getoor, 2005). The co-reference resolution challenge originates from the fact that in a book, the same character will be mentioned using different names, nicknames or even titles. This, together with the fact family members share the same last name or, in some cases, they can have more than one name in common, makes the step of identifying the characters in a fictional novel and the instances where they interact a very difficult task. This is evident not only in NLP tools but also in humans, something made evident by the necessary cautions needed to perform manual annotations in training corpuses (Aroyo & Welty, 2013; Sabou et al., 2014).

The variance of possible mentions that can be used for the same character, combined with the use of pronouns during the novel and anaphoric mentions such as “the doctor” and “the father”, poses many barriers to the automation of this process. Consequentially, most academic works choose to only consider nouns, excluding any nominals and pronouns with the justification that discarding these does not lead to a loss of information (Labatut & Bost, 2019). To achieve this final list, the majority of works employ Named Entity Recognition tools to identify persons present in the text. However as previously established, these do not achieve optimal results in less-spoken languages. As such, a trend as developed where a predefined list of characters, either constituted manually (He et al., 2013) or taken from external sources (Shahsavari et al., 2020; Silva et al., 2023), is matched to the names present in the text. It is crucial to note that this is not suitable for automation since books from less spoken languages may not have this type of data available. A more consistent approach that has been employed and shown promising results in languages other than English, utilizes Part-of-speech tagging (O’Keefe et al., 2012; Rocha et al., 2014). This method takes advantage of other attributes rather than just nouns, taking into account other aspects, for instance gender and honorifics such as “Sir” or “Lady” and double checks for abidance between gender of characters and honorifics (Adanay & Sporleder, 2015) also performing post-processing in the

resulting list, such as redacting names that are outliers or appear less than a pre-determined number of times (Elson et al., 2010).

After extracting the list of characters present in the novel, the second step is detecting interactions. What counts as an interaction changes based on the specific problem that is addressed and there is still debate on what is the most appropriate method to capture the most accurate Social Network (Adanay & Sporleder, 2015).

There are five different approaches to what counts as an interaction between characters in Literary novels. Co-occurrence, where interaction is assigned when two characters appear in a pre-determined unit of text, e.g., a sentence (Adanay & Sporleder, 2015). Conversations, where only verbal interactions are marked as such, this strategy works best for theatre plays since most action is portrayed through speech (Elson & McKeown, 2010; Chak et al., 2017). However, for novels, many interactions between characters can be described in the body of text and thus are lost (Min & Park, 2019), with the advantage being the possibility for directed networks that co-occurrence does not allow for (Moretti, 2011). In some specific cases authors aim not for the standard Social or Conversational Networks that models interactions but instead hope to map the affiliations between the characters present in the Novel (Valls-Vargas et al., 2021), being the nature of the relationship such as enemies or allies (Srivastava et al., 2016) or family relations (Bick, 2023). Another hypothesis that has seen some work devoted to, believes that character interactions can be found in direct actions (Mamede & Chaleira, 2004). In these cases, interactions are attributed to certain verbs that can be found in written text that describe interactions between characters, such as “talked, fought, helped”, and two characters are assigned an interaction when one of these verbs is found between them. The last approach consists of different combinations of the previously mentioned techniques in an attempt to get the most out of each one, such as finding character dialogues and conversations utilizing certain verbs and thus utilizing both conversational and direct actions approaches (He et al., 2010; Agarwal et al., 2013).

Co-occurrence is currently established to be the most appropriate method to detect interactions, it is however relevant to note the theorical challenges it poses. This catch-all method is by nature imprecise, leading to the existence of false positive interactions, however there is proof false negatives are not as common and a case can be made that if two names

co-exist in a sentence, even if it doesn't express a direct interaction between two characters most likely reveals some form of connection between them and as such is not hurtful when mapping out social networks of characters (Adanay & Sporleder, 2015).

The last step of extracting and building character networks is modelling the networks themselves with the acquired data. Although this step is more straightforward, two chains of thought have developed, the most common approach being static networks. In most works character networks are intended to represent the entirety of the novel and for purposes of comparison between different novels. Static networks model character interactions in their totality to draw broader conclusions about the character's social interactions throughout the story (Oelke et al., 2012).

However, some authors have raised the question of how well static networks can capture evolutions and changing relationships between characters throughout a story (Agarwal et al., 2012) and opt instead for dynamic networks. These operate by dividing the narrative into multiple time frames and capturing the social networks from each of these windows, with the most common time unit being the chapters (Adanay & Sporleder, 2015).

3. DATA AND METHODS

This chapter presents the data and methodology employed to conduct this work. Subsection 3.1 defines the corpus used to assess our system's performance and outlines the guidelines for our manual tagging. Subsection 3.2 details the methodology behind our pipeline.

3.1. DATA

We evaluate our Character Network Extraction (CNE) pipeline on a corpus of eleven Portuguese-language novels publicly available at Projecto Adamastor[1]. The corpus consists of six Portuguese and five Brazilian novels from ten different authors, published between 1862 and 1941, as seen in Table 1. This diversity tests our CNE pipeline across various writing styles and is not skewed by overfitting to a specific author's writing (Elson et al., 2010).

Table 1 - List of Novels

Authors	Titles	Year	Nationality
Aluíso Azevedo	O Cortiço	1890	BR
Bernardo Guimarães	A Escrava Isaura	1875	BR
Camilo Castelo Branco	Amor de Perdição A Queda dum Anjo	1862 1866	PT
Eça de Queiroz	O crime do padre amaro	1875	PT
Júlia Lopes de Almeida	A Viúva Simões	1897	BR
Júlio Dinis	As Pupilas do Senhor Reitor	1863	PT
Lima Barreto	O Triste Fim de Policarpo Quaresma	1915	BR
Machado de Assis	Dom Casmurro	1899	BR
Mário de Sá-Carneiro	A Confissão de Lúcio	1914	PT
Soeiro Pereira Gomes	Esteiros	1941	PT

A pre-processing step was performed in the corpus to remove items such as publishing and distribution information and author notes to ensure that only literary text was considered while performing our information extraction process (Dekker et al., 2019).

To benchmark the results of our CNE pipeline, a random chosen chapter from each novel was manually tagged. The tagging process consisted of identifying and flagging every instance of a

[1] https://projectoadamastor.org/

character occurrence in the text, taking in account the possible name variations, thus creating a list of each character and co-referents present in the tagged chapter. Secondly, each time an interaction occurs between two or more characters present in the chapter as seen in Figure 2, creating a dictionary of every interaction between characters present in the chapter.

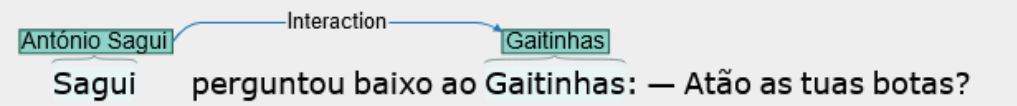


Figure 2- Example of two characters interacting manual annotation

Due to variations in the size of each chapter from book to book, a minimum of 1400 tokens per chapter was established. In case a single chapter would not fulfil the minimum requirement, another would be added to the tagging process. From the 11 manually tagged chapters the smallest one contains 1481 tokens and the largest 5580, in a total of 34904 manually tagged tokens and an average of 3173 tokens per chapter. This benchmark sample is comparable to most work done in the area where resources and time do not allow for a larger annotated sample (Elson et al., 2010; Dekker et al., 2019; Adanay & Sporleder, 2015; Jayannavar et al, 2015).

3.2. METHODS

As established in our Literature Review, an automatic character network extraction system for literary novels needs to perform three tasks (Labatut & Bost, 2019):

1. Extract Character names and perform Co-Reference Resolution
2. Detect interactions between characters
3. Building a graphical representation of results.

Our CNE pipeline follows this standard methodology.

For the first step of extracting character names and performing Co-Reference resolution. As previously mentioned, NER tools in languages other than English tend to underperform in this task as they return a large number of false positives. Therefore, works developed in other

languages adopt the use of Part-of-Speech tagging, which allows for greater supervision and customization, considering the rules specific to each language (Krug et al., 2015).

We follow this POS tagging approach, utilizing two different of-the-shelf tools, spaCy's pt-core-news-lg[2] (Rademaker et al., 2017) and LX-Tagger[3] (Branco & Silva, 2004), to mitigate the out-of-domain use of these tools for the Portuguese language.

Using LX-Tagger, we extract character names by retrieving combinations of tagged words presented in Table 2. These combinations are based on the work done by Trovati & Brady (2014) and Adanay & Sporleder (2015) and are adapted to the rules of Portuguese names:

Table 2- Rules for retrieval of tagged sequences

Tag Pattern	**Example**
Proper Name	Domingos
Title + Proper Name	Sr. Domingos
Proper Name + Proper Name…	Domingos José Correia Botelho
Proper Names + Prepositions/Definite articles if + Proper Name	Domingos José Correia Botelho de Mesquita
Title + Proper Names + Prepositions/Definite articles if + Proper Name	Sr. Domingos José Correia Botelho de Mesquita

The resulting list contains a series of sequences that are not exclusively characters names.

A series of techniques are applied on the resulting list to clean the entries with the goal of achieving a list that contains exclusively character names, and examples of their output are presented in figure 3.

We start by retrieving any sequence that contains a title since we have a guarantee this corresponds to a name. After this, a list of verbs or words that we call presence of a person indicator was constructed: "gritou, suspirou, perguntou, …" (Freitas & Freitas, 2017), we

[2] pt_core_news_lg-3.7.0-py3-none-any.whl
[3] http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXTagger.html#lxtagger

assume that if one of these indicators precedes an identified sequence, it is most likely a character (Agarwal et al., 2012; Trovati & Brady, 2014; He et al., 2013).

The remaining entries on the list are re-tagged by spaCy to eliminate incorrect entries that LX-Tagger may have flagged. We then cross the remaining entries with a geographic database[4] to remove sequences that consist only of cities or locations.

A smaller processing steps is made, consisting of removing incorrectly annexed tokens from the sequence by finding lower-case versions of said tokens in the text. Lastly, the remaining working list is checked with a database of Portuguese names[5]. We end this step with a list of sequences of the various character names and their variations.

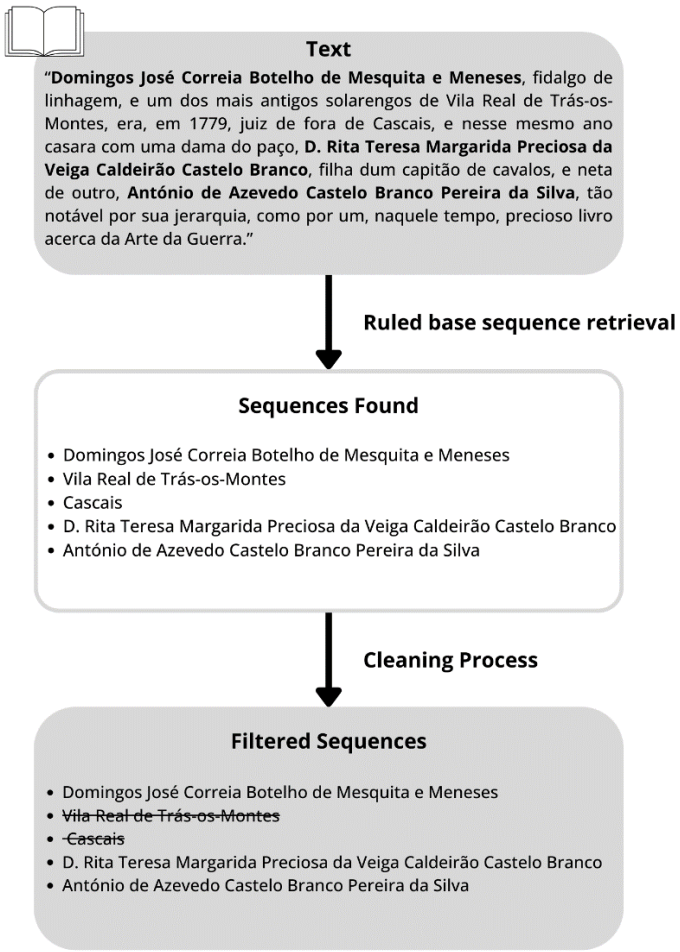


Figure 3- Example of output achieved by each step of sequence retrieval and cleaning process

[4] https://simplemaps.com/data/world-cities

[5] https://irn.justica.gov.pt/Portals/33/Regras%20Nome%20Proprio/Lista%20Nomes%20Pr%C3%B3prios.pdf?ver=WNDmmwiSO3uacofjmNoxEQ%3D%3D

With the resulting list, which is expected to contain only sequences that correspond to character names, we now perform co-reference resolution. Our methodology consists of a set of six heuristics. These were achieved by adapting and combining already existing methodologies (Krug et al., 2015; Elsner, 2012; Vala et al., 2015; Adanay & Sporleder, 2015) to Portuguese names. Examples of the output achieved by each step can be seen inf Figure 4 and they are as follows:

1. We start by iterating through the sequence in descending order from the entry with the highest number of tokens, for the first entry a new character index is created. See Heuristic 1: Entry with most tokens, Figure 4.

2. Each subsequent entry is searched for exact matches in tokens with the already existing character index. If a sequence has an exact match, it is added to the index, if not, a new entry is created. This process is repeated until all multiple token entries have been treated. See Heuristic 2: Matching sequences with multiple tokens, Figure 4.

3. A search is performed in the text for all sequences in a character index. Each time a sequence is found in the text, it is registered and counted. A data frame is constructed with all the counts of each character index throughout the text, and the relevance of each character is determined.

4. Single token entries are then distributed by character index matching according to character relevance and prioritizing first name matching. See Heuristic 3/4: Sorting relevance for single tokens, Figure 4.

5. The names are then checked for matching in a manually constituted list of diminutives and nicknames and turned into their canonical form, if a match is found, the groups that contain the same tokens are joined together. See Heuristic 5: Diminutives and nicknames, Figure 4.

6. Lastly, the system asks us if the book is written in the first person, if so, we are asked to indicate the group corresponding to the narrating character (Gessey-Jones et al., 2020).

This step ends with a list of groups representing one character and all the possible name variations found in the text. See End Product, Figure 4.

Also included in this step but less relevant to our final results is a system that extracts information about character's gender, which was also set up. This information can later be used to develop deeper analysis on the novel and its characters, where gender might be a relevant factor (Elsener, 2012). We achieved this by crossing information with the names present in the database of Portuguese names, the ones not present are put in a voting system where the anteceding tokens to the sequence are retrieved and determined to be male or female based on the most prominent gender. See Gender Voting, Figure 4

At this point, we begin the task of detecting interactions between characters. As referenced in the previous chapter, co-occurrence is, with all its flaws, still the current best solution for interaction detection. Capturing both verbal interactions and interactions that occur in the body of the text and as such the best method when treating novels (Adanay & Sporleder, 2015).

Our CNE pipeline thus considers that an interaction exists when two characters from the list are mentioned in the same text window. This window has been set to 3 sentences (Silva et al., 2023). This frame was determined as the most capable of capturing the majority of interactions without returning a large number of False Positives. A pre-processing step is added where every character that occurs less than three times throughout the novel is eliminated from the list, this helps to clean any previous mistakes made by the system and entries that consist of mentions instead of characters (Elsner, 2012), e.g. "Luís de Camões".

This step returns a table with three columns that present the names of characters who interacted with each other and the number of interactions between them throughout the book, see Interaction Count, figure 4.

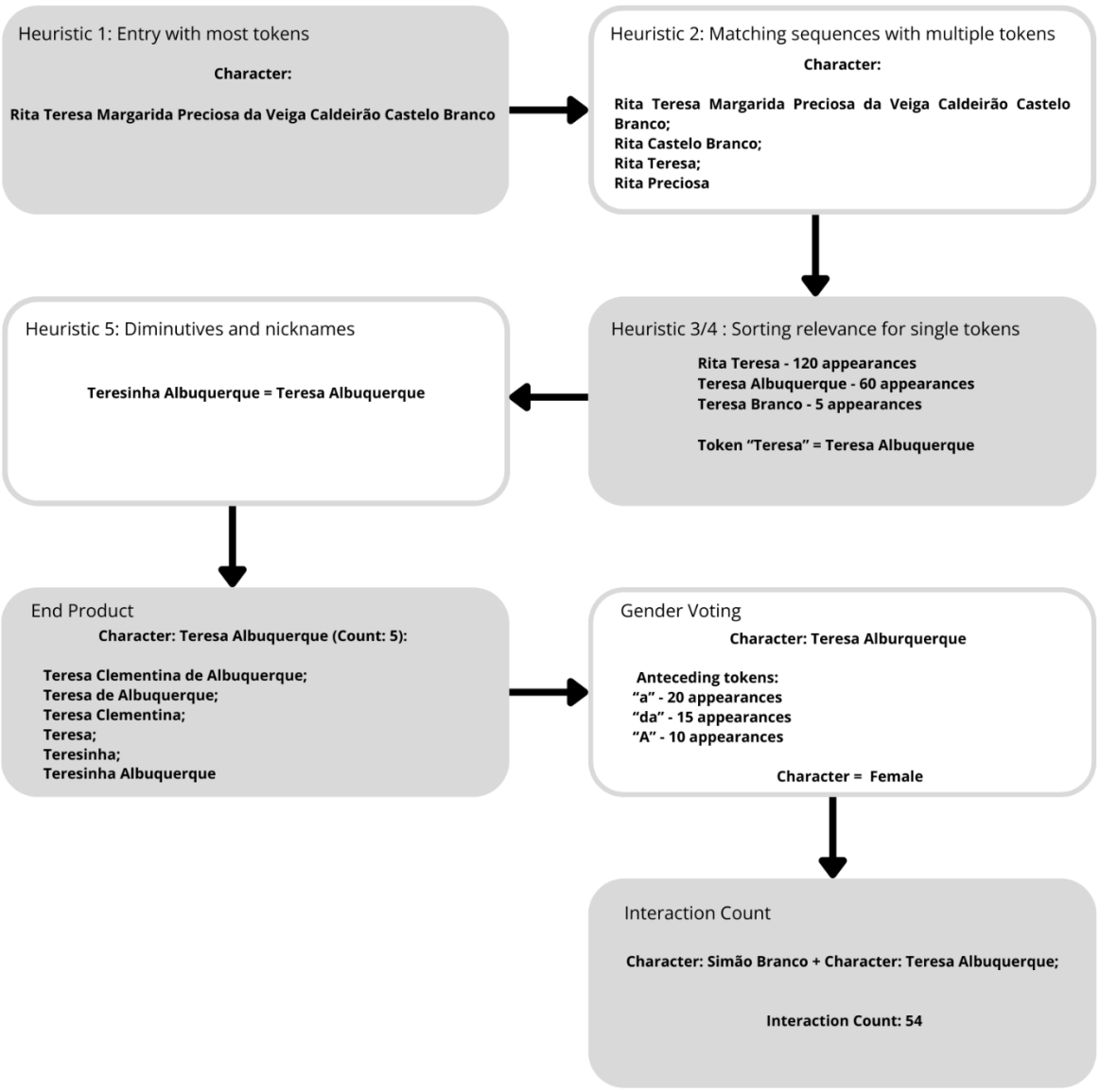
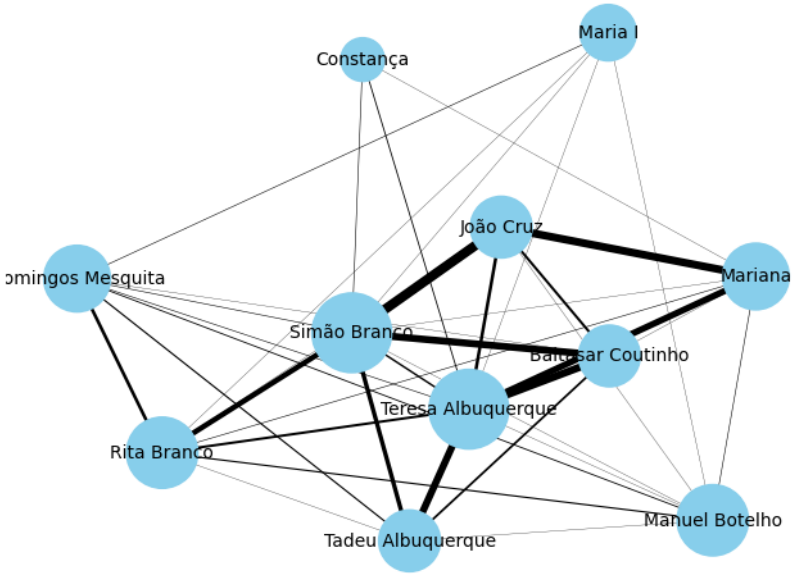


Figure 4- Example of output for each step of Co-reference resolution process

With this, we enter the final step, which is a graphical representation. Using the resulting data frame from our previous two tasks, our system computes and plots a static character network, with characters represented as nodes and interactions as edges, both the nodes and the edges are weighted according to their relevance. Our CNE pipeline then presents three different visualizations for the same network to facilitate analysis. An example of one of this networks can be seen in Figure 3.



.

Figure 5- Character Network from “Amor de Perdição”

4. RESULTS AND DISCUSSION

This work set out to answer how could character network extraction methods be adapted to address the unique linguistic challenges in Portuguese literary works.

To that extend, we developed a CNE pipeline that takes advantage of the available tools for Portuguese language, combining them with previously tested methods, to try and reproduce the results achieved by English language SOTA tools utilized in these tasks.

As already established, we can consider the pipeline is comprised of three subtasks: (I) identifying character occurrences while performing co-reference resolution, (ii) detect interactions between said characters and (iii) the subtask consisting of building a graphical representation of the ensuing results. With the first and second subtasks output being the most significant to our final product.

In order to assess our pipeline's performance in these two subtasks, we not only benchmark the results on the manually tagged chapters but also compare how our pipeline's output to the once achieved by available tools that can perform these tasks in the Portuguese language.

Three metrics will be used in our assessment (Rocha et al., 2014): Precision (1), that answers off all the instances of characters or interactions found in the text by our pipeline, how many of them are actually positive;

$$Precision = \frac{True\ Positives}{True\ Positives + False\ positives} \quad (1)$$

Recall (2), that answers of all the actual positive instances, how many did the pipeline correctly identified;

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

And F1-Score (3), the harmonic mean between both Precision and Recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

We will start with the results of the character occurrence detection process and compare them to those obtained from the off-the-shelf Named Entity Recognition tool spaCy's Portuguese language model "pt_core_news_lg". We do this to have a better understanding of what would be the outcome achieved by a NLP system without all the further steps we implemented. The results are presented in Table 3.

Table 3- Pipeline's metrics in Character Occurrence detection compared to NER

	CNE Pipeline			NER		
List of Novels	Precision	Recall	F1-Score	Precision	Recall	F1-Score
A Confissão de Lucio	**96.6**	**93.4**	**95.0**	*50.0*	*84.6*	*62.9*
A queda de um Anjo	**92.9**	**96.3**	**94.5**	*11.1*	**96.3**	*19.9*
A Viuva Simões	**93.9**	**100**	**96.9**	*43.7*	**100**	*60.8*
Amor de Perdição	**98.3**	**100**	**99.2**	*21.1*	**100**	*34.9*
As Pupilas do Senhor Reitor	**93.7**	**93.7**	**93.7**	*27.1*	**93.7**	*42.0*
Dom Casmurro	**98.4**	**98.4**	**98.4**	*31.2*	*77.4*	*44.4*
Escrava Isaura	**97.8**	*91.7*	**94.6**	*22.6*	**93.8**	*36.4*
Esteiros	**98.8**	*97.7*	**98.2**	*32.0*	**100**	*48.5*
O Cortiço	**90.6**	*77.4*	**83.5**	*36.8*	**90.3**	*52.3*
O Crime do Padre Amaro	**81.3**	**96.8**	**88.3**	*15.1*	*87.2*	*25.7*
Triste Fim de Policarpo Quaresma	**90.8**	**94.7**	**92.7**	*33.2*	*94.7*	*49.2*
Average	**93.9**	**94.5**	**94.1**	*29.4*	*92.5*	*43.4*

Note: The highest achieved score for each metric is highlighted in **bold** and the lowest is highlighted in *italics.*

As can be seen, our system significantly outperforms the off-the-shelf NER both in Precision and F1-Score, with an average of 93.9 and 94.1 against 29.4 and 43.4, respectively. Regarding Recall, the difference is less significant, with our pipeline achieving 94.5 as opposed to 92.5

for the NER tool. We can thus conclude that the NER tool is able to correctly identify instances of characters in novels but does not do so without returning a large volume of False Positives. This is to be expected and explained by all the previously mentioned traits of literary text that hinder NLP capabilities and performance.

Overall, our Pipeline achieved very promising results in this task. With no other systems developed for the Portuguese language to compare to, our scores indicate the pipeline was very competent in identifying character names and their variations in the manually tagged samples. The lowest performances, seen for “O Cortiço” with 77.4 Recall, can be correlated to the sample chapter itself, where a total of 27 characters were manually identified, with 15 making one single appearance in the text. This indicates that our pipeline might struggle with identifying minor characters.

As per the lowest Precision recorded for “O Crime do Padre Amaro” with 81.3. This sample chapter had the largest number of tokens, with a Recall of 96.8, our system retrieved almost the entirety of the 94 total instances of characters with their corresponding name variations. However, it returned 20 false positives. To this extent, we can assess that our system when performing this task in a full novel as intended, is very capable of returning the correct instances of characters solving for co-reference but is unable to do so without returning a marginal number of misclassified entries.

Regarding the second subtask of correctly identifying interactions between characters in a novel, we will compare our systems performance with ChatGPT 4.0. This choice was made to compare our CNE pipeline’s performance against a readily available tool that can perform this task and test the current capabilities of Artificial Intelligence to further develop this field. This Large Language Model was given an example of a character interaction data frame and was asked to replicate this upon being fed each chapter under analysis, a manual step was performed in the results regarding character’s name variations, since this task was meant to test capabilities on interaction detection and not co-reference resolution. The results are as seen in Table 4.

Table 4- Pipeline's metrics in detected interactions compared to ChatGPT

	CNE Pipeline			ChatGPT		
List of Novels	Precision	Recall	F1-Score	Precision	Recall	F1-Score
A Confissão de Lucio	**80.0**	**88.9**	**84.2**	*11.4*	*55.6*	*18.9*
A queda de um Anjo	*55.6*	**100**	**94.5**	**100**	*40.0*	*57.1*
A Viuva Simões	**66.7**	**100**	**80.0**	**66.7**	*75.0*	*70.6*
Amor de Perdição	**81.3**	**81.3**	**81.3**	*41.7*	*62.5*	*50.0*
As Pupilas do Senhor Reitor	*75.9*	**91.7**	**83.0**	**83.3**	*62.5*	*71.4*
Dom Casmurro	**56.0**	**87.5**	**68.3**	*40.6*	*81.3*	*54.2*
Escrava Isaura	*57.1*	**100**	*72.7*	**90.0**	*75.0*	**81.8**
Esteiros	**73.7**	**73.7**	**73.7**	*46.4*	*68.4*	*55.3*
O Cortiço	**52.4**	**64.7**	**57.9**	*27.5*	**64.7**	*38.6*
O Crime do Padre Amaro	*48.9*	**95.8**	**64.8**	**100**	*25.0*	*40.0*
Triste Fim de Policarpo Quaresma	**69.2**	**87.1**	**77.1**	*68.4*	*41.9*	*52.0*
Average	**65.1**	**88.2**	**76.1**	*61.5*	*59.3*	*53.6*

Note: The highest achieved score for each metric is highlighted in **bold** and the lowest is highlighted in *italics.*

Our system once again outperformed the results obtained from its counterpart. ChatGPT struggles with larger chapters, indicating that it would be unable to perform this task in a full novel. Its performance seems to be highly dependent on the writing style, sometimes presenting high Precision but doing so at the cost of a lower Recall, as we can see in "O Crime do Padre Amaro" with 100 and 25.0 respectively, or presenting an acceptable Recall, but doing so with low Precision as in "O Cortiço", with 64.7 and 27.5 respectively.

Our pipelines assessment has shown itself to be more consistent. With an average of 88.2 Recall and 65.1 Precision, our system seems to be proficient in correctly identifying instances of characters interacting but does so with a few false positives.

Taking an in-depth look at this issue in "A queda de um Anjo", where Precision was 55.6 and Recall 100, we see that only five interactions were manually identified in the chapter, all were correctly identified. Our system, however, identified four more interactions, all of which corresponding to characters that were present in this chapter.

The same can be said for “A Escrava Isaura”, where, once again, there were no false negatives out of the twelve manually tagged interactions. Our system, again, mistakenly recorded more interactions between the characters present in the text and one instance of a character who was not present. This can be explained by the mention of “S. António”, a religious sanctity that our system incorrectly identified as the character “António”. The rest of the false positives were interactions between characters present in the chapter that were not tagged as such manually.

Overall, our pipeline achieved positive results in both the assessed tasks and is an improvement from the currently available tools that can perform these tasks as seen in Table 5. The mistakes we see from the pipeline can be attributed to the ambiguity of what counts as an interaction, for example, two characters exchange back forth in a dialogue, in this case, our pipeline might find more interactions than the ones manually assessed. Seeing our system’s final goal is to represent graphically a network of characters and how connected they are throughout a novel, this is less significant.

Table 5 – Average F1-Score and F1-Score increase

Task	Average F1-Score	F1-Score increase from available tools
Identifying character and co-reference	94.1	50.7
Detect interactions	75.9	22.3

The same can be said for the lower scores in chapters where few interactions occurred. As in the previously mentioned “A queda de um Anjo” chapter, having a low total number of manually tagged interactions makes one mistake by the pipeline very costly on the Precision score. However, when we think of the end goal of our pipeline, one incorrectly tagged interaction is not as relevant in the end product (Adanay & Sporleder, 2015).

Other identified problems, such as religious evocations incorrectly identified as characters, are unfortunately a necessary trade-off. In some books, relevant character might be identified the same way as saints, such as “S. Joaneira” in “Crime do Padre Amaro. This serves as a reminder of how complex the task at end is.

5. CONCLUSIONS

We built a pipeline that automates the Character Network extraction process and can be applied to a variety of novels. It takes advantage of techniques employed in other languages and improves on results obtained using off-the-shelf and state-of-the-art tools.

By reproducing methodologies tested in other languages and adapting them to the Portuguese language rules, we overcame the difficulty NLP tools present while trying to perform this task in Portuguese novels. To our knowledge, this is the first automated Character Network Extraction system for the Portuguese language that can be applied, with no requirement for external annotations, to a vast corpus of novels.

Through this work, we have tested and advanced the capabilities of NLP in processing complex texts in Portuguese Language. By deepening our understanding of these text mining tools and enhancing them, we aim to further develop text mining capabilities in Portuguese, that can be utilized for digital marketing purposes, helping us better understand market trends and customer sentiments.

Overall, our pipeline has achieved very promising results, with a 94.1 average F1-Score for identifying character occurrences considering co-reference and a 75.9 F1-Score for detecting interactions between said characters. These results suggest that future work on our methods can be performed. Anaphoric resolution for co-reference is still a challenge even for English models with access to more effective tools, some testing was performed in our pipeline to address this, but no satisfying results have been achieved. Another aspect that can be improved is co-occurrence as an interaction detection method. Its imprecise nature is known within the field, but it is still accepted as the best available method. Future exploration in this matter is needed.

It is important to note that the lack of an available manually annotated corpus limited our ability to evaluate our system's performance. Although blindly choosing chapters from a corpus of books is standard practice, our system has yet to be tested on a full set of novels, limiting our conclusions. It is also necessary to note that the availability of freely accessible books limited our corpus. This is a common problem and is why most works developed in Character Network Extraction are conducted in 19th-century literature. As such our pipeline

has yet to be tested with more modern and contemporary books, which can have different aspects to consider when solving for co-reference. However, the novel “Esteiros” published in 1941, has all the characteristics of modern books regarding character names, since most characters are treated by their nicknames as opposed to full names, and our pipeline performed positively to this test.

To this extent, our pipeline will be made publicly available[6], not only to allow for studies in Character Networks Analysis but also to encourage further exploration of these issues and progress in work done for the Portuguese language.

[6] https://github.com/ticadu/taggus

BIBLIOGRAPHICAL REFERENCES

Alberich, R., Miro-Julia, J., & Rossello, F. (2002). Marvel universe looks almost like a real social network. *arXiv/0202174.* https://arxiv.org/abs/cond-mat/0202174

Agarwal, A., Corvalan, A., Jensen, J., & Rambow, O. (2012, June 1). Social network analysis of *Alice in Wonderland*. *ACLWeb*; Association for Computational Linguistics. https://aclanthology.org/W12-25

Agarwal, A., Kotalwar, A., Zheng, J., & Rambow, O. (2013, October 1). SINNET: Social interaction network extractor from text (K. Torisawa & H. Li, Eds.). *ACLWeb*; Asian Federation of Natural Language Processing. https://aclanthology.org/I13-2009

Aroyo, L., & Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. https://www.semanticscholar.org/paper/Crowd-Truth%3AHarnessing-disagreement-in-a-relation-AroyoWelty/672d4ffb4895070da74a22b027a215a7adbebb9e

Bhattacharya, I., & Getoor, L. (2005). Relational clustering for multi-type entity resolution. https://doi.org/10.1145/1090193.1090195

Bick, E. (2023). Extraction of literary character information in Portuguese. *Linguamática, 15*(1), 31–40. https://doi.org/10.21814/lm.15.1.397

Bolioli, A., Casu, M., Lana, M., & Roda, R. (2013). Exploring the betrothed lovers. *OASIS* 30–35. DOI: 10.4230/OASIcs.CMN.2013.30

Bossaert, G., & Meidert, N. (2013). "We are only as strong as we are united, as weak as we are divided": A dynamic analysis of the peer support networks in the *Harry Potter* books. *Open Journal of Applied Sciences, 3*(2), 174–185. https://doi.org/10.4236/ojapps.2013.32024

Bornet, C., & Kaplan, F. (2017). A simple set of rules for character and place recognition in French novels. *Frontiers in Digital Humanities, 4.* https://doi.org/10.3389/fdigh.2017.00006

Branco, A., & Silva, J. (2004, May 1). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese (M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva,

Eds.). *ACLWeb*; European Language Resources Association (ELRA). https://aclanthology.org/L04-1354/

Callison-Burch, C., & Dredze, M. (2010, June 1). Creating speech and language data with Amazon's Mechanical Turk (C. Callison-Burch & M. Dredze, Eds.). *ACLWeb*; Association for Computational Linguistics. https://aclanthology.org/W10-0701/

Chak, Y., Yeung, J., & Lee, S. (2017). Identifying speakers and listeners of quoted speech in literary works (pp. 325–329). https://aclanthology.org/I17-2055.pdf

Choi, Y.-M., & Kim, H.-J. (2007). A directed network of Greek and Roman mythology. *Physica A: Statistical Mechanics and Its Applications, 382*(2), 665–671. https://doi.org/10.1016/j.physa.2007.04.035

Coll Adanay, M., & Sporleder, C. (2015). Clustering of novels represented as social networks. *Linguistic Issues in Language Technology, 12.* https://aclanthology.org/2015.lilt-12

da Silva Conrado, M., Felippo, A. D., Salgueiro Pardo, T. A., & Rezende, S. O. (2014). A survey of automatic term extraction for Brazilian Portuguese. *Journal of the Brazilian Computer Society, 20*(1). https://doi.org/10.1007/s13173-014-0118-0

Dekker, N., Kuhn, T., & van Erp, M. (2019). Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science, 5.* 10.7717/PEERJ-CS.189

Elson, D. K., & McKeown, K. (2010). Automatic attribution of quoted speech in literary narrative. *Proceedings of the AAAI Conference on Artificial Intelligence, 24*(1), 1013–1019. https://doi.org/10.1609/aaai.v24i1.7720

Elson, D. K., Dames, N., & McKeown, K. (2010). Extracting social networks from literary fiction. *Meeting of the Association for Computational Linguistics,* 138–147. https://aclanthology.org/P10-1015

Elsner, M. (2012). Character-based kernels for novelistic plot structure (W. Daelemans, Ed.). *ACLWeb*; Association for Computational Linguistics. https://aclanthology.org/E12-1065

Fonseca, B. P., Albuquerque, P. C., Zicker, F., & Morel, C. M. (2021, July 18). Social network analysis and mining: Challenges and applications. *Sol.sbc.org.br*; SBC. https://doi.org/10.5753/brasnam.2021.16149

Freitas, B., & Freitas, C. (2017). Verbos de elocução em português: um estudo descritivo com base em grandes corpora e motivado pela linguística computacional. *Fórum Lingüístico, 14*(3), 2266–2266. https://doi.org/10.5007/1984-8412.2017v14n3p2266

Gessey-Jones, T., Connaughton, C., Dunbar, R., Kenna, R., MacCarron, P., O'Conchobhair, C., & Yose, J. (2020). Narrative structure of *A Song of Ice and Fire* creates a fictional world with realistic measures of social complexity. *Proceedings of the National Academy of Sciences, 117*(46), 28582–28588. https://doi.org/10.1073/pnas.2006465117

Grayson, S., Wade, K., Meaney, G., & Greene, D. (2016). The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature. *IFIP Advances in Information and Communication Technology,* 65–77. https://doi.org/10.1007/978-3-319-46224-0_7

He, H., Barbosa, D., & Kondrak, G. (2013). Identification of speakers in novels. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.*

He, H., Kondrak, G., & Barbosa, D. (2010). The actor-topic model for extracting social networks in literary narrative. DOI: 10.3115/1679044.1679045

Jayannavar, P., Agarwal, A., Ju, M., & Rambow, O. (2015, June 1). Validating literary theories using automatic social network extraction (A. Feldman, A. Kazantseva, S. Szpakowicz, & C. Koolen, Eds.). *ACLWeb*; Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0704

Jha, M., Andreas, J., Thadani, K., Rosenthal, S., & McKeown, K. (2010, June 1). Corpus creation for new genres: A crowdsourced approach to PP attachment (C. Callison-Burch & M. Dredze, Eds.). *ACLWeb*; Association for Computational Linguistics. https://aclanthology.org/W10-0702/

Jha, M., Andreas, J., Thadani, K., Rosenthal, S., & McKeown, K. (2010, June 1). *Corpus creation for new genres: A crowdsourced approach to PP attachment* (C. Callison-Burch & M. Dredze,

Eds.). ACLWeb; Association for Computational Linguistics. https://aclanthology.org/W10-0702/

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). *Text mining for market prediction: A systematic review*. Expert Systems with Applications, 41(16), 7653–7670. https://doi.org/10.1016/j.eswa.2014.06.009

Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., & Weimar, L. (2015, June 1). *Rule-based coreference resolution in German historic novels* (A. Feldman, A. Kazantseva, S. Szpakowicz, & C. Koolen, Eds.). ACLWeb; Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0711

Labatut, V., & Bost, X. (2019). *Extraction and analysis of fictional character networks*. ACM Computing Surveys, 52(5), 1–40. https://doi.org/10.1145/3344548

Lee, J., & Wong, T. (2016). *Conversational network in the Chinese Buddhist Canon*. Open Linguistics, 2(1). DOI:10.1515/opli-2016-0022

Mac Carron, P., & Kenna, R. (2012). *Universal properties of mythological networks*. EPL, 99(2), 28002. https://doi.org/10.1209/0295-5075/99/28002

Mamede, N. J., & Chaleira, P. (2004). *Character identification in children stories*. Lecture Notes in Computer Science, 82–90. https://doi.org/10.1007/978-3-540-30228-5_8

Trovati, M., & Brady, J. P. (2014). *Towards an automated approach to extract and compare fictional networks: An initial evaluation*. https://doi.org/10.1109/dexa.2014.58

Trovati, M., Bessis, N., Huber, A., Zelenkauskaite, A., & Asimakopoulou, E. (2014). *Extraction, identification, and ranking of network structures from data sets*. https://doi.org/10.1109/cisis.2014.46

Min, S., & Park, J. (2019). *Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling*. PLoS ONE, 14(12). https://doi.org/10.1371/journal.pone.0226025

Moretti, F. (2011). *Literary Lab network theory, plot analysis*. https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf

Oelke, D., Kokkinakis, D., & Malm, M. (2012, April 1). *Advanced visual analytics methods for literature analysis* (K. Zervanou & A. van den Bosch, Eds.). ACLWeb; Association for Computational Linguistics. https://aclanthology.org/W12-1007/

Mac Carron, P., & Kenna, R. (2012). *Universal properties of mythological networks*. EPL, 99(2), 28002. https://doi.org/10.1209/0295-5075/99/28002

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., & de Paiva, V. (2017, September 1). *Universal dependencies for Portuguese*. ACLWeb; Linköping University Electronic Press. https://aclanthology.org/W17-6523/

Rieck, B., Leitte, H., Alonso, A., Boatswain, A., Ceres, C., Gonzalo, F., Juno, I., Prospero, M., Stephano, S., Adrian, T., Antonio, A., Boatswain, A., & Sebastian, M. (2016). *"Shall I compare thee to a network?" Visualizing the topological structure of Shakespeare's plays*. Retrieved November 22, 2023, from https://bastian.rieck.me/research/Vis2016.pdf

Rocha, M., Jorge, A., Oliveira, M., Brito, P., Gama, J., & Pimenta, C. (2014). *From entity extraction to network analysis: A method and an application to a Portuguese textual source*. DOI: 10.1145/268391210.1145/2683912

Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). *Corpus annotation through crowdsourcing: Towards best practice guidelines*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2435b48cc30e30f3849b9670f263b501ba9c0024

Sack, G. (2021). *Character networks for narrative generation*. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 8(2), 38–43. https://doi.org/10.1609/aiide.v8i2.12541

Saura, J. R. (2020). *Using data sciences in digital marketing: Framework, methods, and performance metrics*. Journal of Innovation & Knowledge, 6(2), 92–102. https://doi.org/10.1016/j.jik.2020.08.001

Shahsavari, S., Ebrahimzadeh, E., Shahbazi, B., Falahi, M., Holur, P., Bandari, R., Tangherlini, T. R., & Roychowdhury, V. P. (2020). *An automated pipeline for character and relationship extraction from readers*. https://doi.org/10.1145/3394231.3397918

Silva, M. O., Oliveira, G. P., & Moro, M. M. (2023, August 6). *Analyzing character networks in Portuguese-language literary works*. Sol.sbc.org.br; SBC. https://doi.org/10.5753/brasnam.2023.230585

Srivastava, S., Chaturvedi, S., & Mitchell, T. (2016). *Inferring interpersonal relations in narrative summaries*. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.10349

Vala, H., Jurgens, D., Piper, A., & Ruths, D. (2015). *Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 769–774. Association for Computational Linguistics. https://aclanthology.org/D15-1088

Valls-Vargas, J., Zhu, J., & Ontañón, S. (2021). *Toward automatic role identification in unannotated folk tales*. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 10(1), 188-194. https://doi.org/10.1609/aiide.v10i1.12732

Yuen, M.-C., King, I., & Leung, K.-S. (2011). *A survey of crowdsourcing systems*. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. https://doi.org/10.1109/passat/socialcom.2011.203

Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). *Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing*. Journal of Medical Internet Research, 15(4), e73. https://doi.org/10.2196/jmir.2426

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa