v. 34, n. 1, p. 130–153, 2026
ISSN: 2237-2083 · DOI: 10.17851/2237-2083.34.1.130-153
Submetido em: 22/10/2024 · Aprovado em: 04/12/2025

REVISTA DE ESTUDOS DA LINGUAGEM

Building the Terminology of an Audiovisual Saga: The Example of Star Wars[1]

Elaborando a terminologia de uma saga audiovisual: o exemplo de Star Wars

Guilherme Fromm
Universidade Federal de Uberlândia (UFU) | Uberlândia | MG | BR
guifromm@ufu.br
https://orcid.org/0000-0001-5654-0135

Guilherme Rodrigues Ferreira
Universidade Federal de Uberlândia (UFU) | Uberlândia | MG | BR
guilherme.ferreira1@ufu.br
https://orcid.org/0000-0003-2030-2594

Abstract: The objective of this article is to show how a multimedia franchise can be studied through terminological lenses. The Star Wars saga was chosen for the study due to its huge amount of material available on the internet. The theoretical framework comprises the Terminology area (more specifically, the Communicative Theory of Terminology and Ethnoterminology), the Terminography, a lexical analysis suite (WordSmith Tools) used for the work with concordance lines from the saga and the development of the Terminology in Fiction platform. The methodology presents many steps, from the choice of the material, the selection of term candidates from the WordSmith Tools suite up to the exploration of the concordance lines of these candidates to create entries in the Terminology in Fiction platform. The platform is also analysed, from the creation of the project up to the entries development. As a result, we present one example (of many available at the platform) of a Star Wars monolingual entries in contrast (available in Portuguese and English), useful for translators, fans and for those curious about how terminology permeates even fictional works.

Keywords: terminology; terminography; Star Wars; corpus linguistics; terminology in fiction.

Resumo: O objetivo deste artigo é mostrar como uma franquia multimídia pode ser estudada por meio de lentes terminológicas. A saga Star Wars foi escolhida para o estudo devido à sua enorme quantidade de material

[1] The description of trademarks owned by Disney were used in this paper only for scientific purposes under Brazilian law.

disponível na internet. O arcabouço teórico compreende a área de Terminologia (mais especificamente, a Teoria Comunicativa da Terminologia e a Etnoterminologia), a Terminografia, uma suíte de análise lexical (WordSmith Tools) usada para o trabalho com linhas de concordância da saga e o desenvolvimento da plataforma Terminologia em Ficção. A metodologia apresenta muitas etapas, desde a escolha do material, a seleção de candidatos a termos através da suíte WordSmith Tools até a exploração das linhas de concordância desses candidatos para criar verbetes na plataforma Terminologia em Ficção. A plataforma também é analisada, desde a criação do projeto até o desenvolvimento dos verbetes. Como resultado, apresentamos um exemplo (de muitos disponíveis na plataforma) de verbetes monolíngues de Star Wars em contraste (português e inglês), útil para tradutores, fãs e para aqueles curiosos sobre como a terminologia permeia até mesmo obras de ficção.

Palavras-chave: terminologia; terminografia; Star Wars; linguística de *corpus*; terminologia em ficção.

1 Introduction

Star Wars is one of the most important and influential franchises in the history of cinema, television, books, video games and comics. Since the release of the first film in 1977, the saga created by George Lucas has become a phenomenon and a valuable brand, now owned by The Walt Disney Company. Among the many reasons that motivated the completion of this corpus-based research on Star Wars, the influence of the saga was one of the most significant. It is a large multimedia franchise that has been continually expanding since the 1970s, requiring daily teams of audiovisual translators who must tackle the challenging standardization required for such a vast universe. Currently, the high demands of dubbing and subtitling companies have created a truly frenetic pace, with extremely tight deadlines and increasingly larger projects. Among movies, animations, and TV series, the Star Wars franchise currently has dozens of titles in various stages of production, which will eventually be completed and sent to these companies for their respective translations and adaptations. Considering the work of dubbing and subtitling translators, who deal with complex deadlines and may not necessarily have a deep understanding of fictional universes like that one of Star Wars, terminology assistance is naturally welcomed and motivates the development of this research.

Every aspect of Star Wars presents a vast terminological universe that can be explored. The geography of the saga specifies various planets, terrains, and regions, each with their own names and characteristics. Politics, in turn, involves factions and organizations such as the

Republic, the Rebel Alliance, the Empire, and the First Order, with each institution having a range of affiliations, including military ranks, technologies used, and cultural and economic elements. The franchise also delves into different races, beliefs, philosophies, institutions, and technologies, with each of these divisions carrying plenty of terms (fictional or not) that can be studied in literature or applied in the professional and amateur world of audiovisual translation, given the constant releases in cinema, TV, video games, and literature.

1.1 Goals

The general objective of the project developed by one of the authors TCC[2] (Ferreira, 2022) was to create a vocabulary proposal about the Star Wars franchise on the Terminology in Fiction platform (version 2; query page available at: http://ic.votec.ileel.ufu.br/?lang=en; choose an Area > Sciences and Fiction > Sagas > Star Wars; click the  icon). As for the platform, the intention of this work is to help translation apprentices and professional translators with their jobs. It is expected that, with the implementation of this proposal, the work of translators who deal with the franchise might be improved (eliminating or decreasing the possibility of multiple translations of the same term) and facilitated (from the terminology search interface of the Terminology in Fiction platform).

The TCC specific objectives were to: a. Find candidate terms in English used in subtitles and books; b. Find candidate terms in Portuguese used by translators; c. Analyze the most used vocabulary in English and its Portuguese counterpart in order to reflect on issues of translation; d. Make the online vocabulary proposal available for free through the Terminology in Fiction platform (TF, from now on; Fromm, 2024).

Therefore, this paper assumes that there is an expressive community of translators who can benefit from including Star Wars vocabulary in TF project. In the next section, some of the elements of Star Wars will be addressed, as well as translation decisions that the franchise has gone through around the world over the years.

2 About Star Wars

When released in theaters in 1977, George Lucas's Star Wars was an instantaneous phenomenon. At the time, two sequels were released (The Empire Strikes Back, 1980, and The Return of the Jedi, 1983) forming what is now known as the classic trilogy. One second trilogy was released between 1999 and 2005 and a third one between 2015 and 2019. These films make up the Skywalker Saga, which tells the story of a family of people sensitive to The Force, fighting for the fate of a very distant galaxy (as they point out at the beginning of each film), from the rise, fall and redemption of Anakin Skywalker, the adventures of his children (Luke and Leia), to the young Rey's journey to find her place in the universe. The franchise also has movies independent to the saga, among animations (Star Wars: The Clone Wars, 2008), films produced for television (The Star Wars Holiday Special, 1978; Caravan of Courage: An Ewok Adventure, 1984 and Ewoks - The Battle of Endor, 1985) and spin-offs (Rogue One: A Star Wars Story, 2016

[2] Trabalho de Conclusão de Curso – undergratuate thesis.

and Han Solo: A Star Wars Story, 2018). Released simply as Star Wars, it was only with the success of VHS that the films were finally renamed along the lines of the original vision of author George Lucas. The first film then becomes "Star Wars: Episode IV – A New Hope", indicating that the future would bring a new trilogy about the past of this universe.

According to Lucas (2011), the film was not expected to become such a success, since several elements of the project were exactly opposite to the conventions of science fiction cinema and space operas at the time. Inspired by Akira Kurosawa, the first act of A New Hope establishes who are some of the plot's most powerful and relevant characters, but the story is defined from the point of view of the two figures of minor importance. In Episode 4, these characters are the droids C-3PO and R2-D2, while Kurosawa makes the same dynamic with Tahei and Matashichi, in The Hidden Fortress.

The mythology of Star Wars has been built, since the first film, around real-life elements as politics, religion, and the polarization of good versus evil. Religion, for example, is one of the central themes in the saga, being represented by the Force, a kind of spiritual energy that connects all living elements in the universe. It is then seen as a religion for the Jedi, a group that seeks balance in the galaxy by following a strict moral code linked to ethical and philosophical principles. Politics, on the other hand, plays an important role in the saga, whose narrative is developed in a galaxy with political and ideological fractures, in which characters constantly compete for dominance of power. Topics as tyranny, democracy, corruption, and political oppression, for instance, are addressed throughout of the trilogies involving complex political issues, such as corruption in the Senate and the tyranny of the Galactic Empire.

From the grandeur of the universe established in the classic trilogy, a huge amount of additional material has been produced since the release of the first film in 1977. Among novels, comics, encyclopedias, games and, more recently, several television series, the Star Wars universe gained an established mythology in hundreds of copies in various media, constituting a vast universe that encompasses a significant number of planets, vehicles, groups, technologies, weapons, and droids, for example.

With the constant expansion of the Star Wars universe, the franchise very soon became a brand, present worldwide by becoming a symbol of pop culture. An example of the efforts, so far by George Lucas and the former 20th Century Fox (now a division of the Disney Group), to strengthen the marketing character of the franchise, is evident in the way in which it becomes recognized worldwide with the end of the first trilogy. The films, which until then had their titles translated around the world (La guerre des étoiles, in France; Guerre stelari, in Italy and Guerra nas Estrelas, in Brazil, for example), became known worldwide simply as Star Wars with the release of the second trilogy, in 1999.

Since the release of the first film, it is safe to say that the franchise has secured an important space in popular culture. In 2015, with the release of Star Wars: Episode VII – The Force Awakens, the franchise was officially revived and currently lives its most active moment, with several series, animated and in live-action, in activity and being translated, through dubbing or subtitling, around the world.

Therefore, it is important to highlight the complexity of these translation works in the more diverse media. Due to the short deadlines required by agencies, studios and companies requesting translation and localization work, the professional may encounter difficulties in locating the specifics of the technical vocabulary of franchises like Star Wars. Such difficulties facilitate the incidence of errors, since a shorter research time might result in terminological mistakes.

3 Theoretical framework

In this section, the theoretical references that support the research will be analyzed. We will start from the studies of Terminology (more specifically from the Communicative Theory of Terminology, by Cabré, 1999), Terminography, the studies of Barbosa's Ethnoterminology (2006, extending its principles from literature to audiovisual), and the resources used – WordSmith Tools 7, and the TF project.

3.1 Terminology

It is called "Terminology" (with the initial capital letter) the study about the terms and "terminology" (with the initial lowercase letter) referring to the specific terms and vocabularies of a certain area, being an interdisciplinary field that offers support to several other subjects. According to Lara (2004), "Terminology" has a polysemic meaning, encompassing both theoretical and methodological approaches as well as concrete terminology. A theoretical perspective offers methodologies that describe, organize and transfer knowledge, establishing principles that govern the compilation, formation of terms, structuring of conceptual fields, and the use and administration of terminologies. The meaning of concrete terminology, however, refers to a set of terms related to a specialized language (Lara, p. 234).

Still according to Lara (2005), the Terminological Unit is the object of Terminology. Unlike a word, which encompasses various possibilities of language use, a term refers to words used in specific communication situations, meaning language in specialized contexts. She also establishes that Terminology has both theoretical and practical objectives. The theoretical approach emphasizes the position of terminologies in the knowledge and practice of experts. The goal is to analyze, through specific methods, the use of terminologies in different contexts. The practical approach, on the other hand, occurs, for example, in areas such as "translation, documentation, linguistic standardization, and linguistic planning" (Lara, 2005, p. 6). It is safe to say, then, that it plays a fundamental role in the development of linguistic resources as vocabularies and glossaries, for instance.

Terminological work has seen an increasing use of technological tools developed within the scope of Computational Linguistics and has been increasingly reliant on Corpus Linguistics methodology. The goal, according to Zamora (2015), is to provide a precise description of terminology in a specific field of knowledge, establishing linguistic or terminological equivalences more accurately for research focused on translation (Zamora, 2015). In the realm of literary translation, video game translation, and audiovisual content, such studies represent a potential advancement for the quality of various translated materials. When dealing with expanded universes and well-established sagas, it is quite common to encounter an extensive terminological field which, without proper attention, might result in different translations of the same term as new products are released and translated by different individuals and teams. Terminological analysis through a corpus can assess historically established linguistic patterns by frequency, thereby avoiding or eliminating the possibility of multiple translations for the same term.

Thus, it can be concluded that Terminology plays a fundamental role in the study of specific vocabulary both in real or fictional languages/universes, and by understanding the importance of the field, it is possible to explore its application in many contexts. The study of a particular terminology ensures an understanding of the nature of terms and their uses, which aids in standardization and optimization of work across various fields. In summary, Terminology is, according to Dias (2000, p. 90), in practice, "a set of methods and activities aimed at collecting, describing, processing, and presenting terms; as a product, it is a set of terms, or vocabulary, of a specific specialty".[3]

3.2 Communicative Theory of Terminology

Upon analyzing the foundations of Maria Teresa Cabré works, Fromm (2020) observes that the Communicative Theory of Terminology presents itself in an antagonistic manner to the General Theory of Terminology of Wüster. The latter, rooted in a considerably rigid notion regarding term analysis, showed little flexibility in the study of contexts. Thus, the distinctive aspect of the Communicative Theory of Terminology lies in its emphasis on the dimension of how a term can function, varying between common language and specialized areas; this means that any word can be a term in a specific specialized context (in a vocabulary or glossary), while being merely a lexeme (in a dictionary) in other more general linguistic contexts.

Consequently, Cabré (1999) establishes the Communicative Theory of Terminology to recognize the interdisciplinary nature of Terminology, the relationship between general and specialized knowledge, and the interdisciplinarity and plurality of terms. Additionally, the theory values the polysemy of lexical units, the existence of synonyms, and the diversification of discourse concerning the topic.

3.3 Ethnoterminology

According to Latorre, Ethnoterminology consists of the "intersection zone between the studies of literary discourses and those of specialized languages or terminologies" (2013, p. 73). This work uses the concept of Ethnoterminology, widely applied in the literary field, to analyze technical discourse in a multimedia universe of scientific fantasy.

The Star Wars universe is brimming with elements existing in real life but reimagined in its mythology to establish a rich world composition anchored in bases familiar to the audience. An example is the concept of the "Force" itself, which gains religious values when applied in the narrative (as can be seen in the definition present in TF). This movement is reflected in terminological analysis when terms are encountered that, within the established fictional universe, take on new meanings and/or broader or specific definitions. Regarding the connections between common language and specialized languages, Barbosa warns that "[...] lexical units are multifunctional. The precise establishment of their function depends on their insertion into a discursive norm, which then determines the status of word or term" (2006).

[3] Originally, in Portuguese: "[...] um conjunto de métodos e atividades voltado para coleta, descrição, processamento e apresentação de termos; como produto, é um conjunto de termos, ou vocabulário, de uma determinada especialidade."

According to Fromm (2011), the pursuit of verisimilitude in narratives results in the presence of specific terminology in fictional works. To reach this conclusion, one of Fromm's advisees analyzed Star Trek (Peixoto, 2014), a series relatively similar to our studies (also a space saga) and observed the massive presence of unique terminology derived from real fields such as Physics, Astronautics, and Astronomy. Similar trends can be found in Star Wars, in the fields of Physics (Hyperspace, Lightspeed) and Militarism (Star Destroyer), for example.

Therefore, this paper is based on the premise that the same approach can be applied to analyzing the Star Wars universe, considering the rich world-building established by various authors since the 1970s through comics, books, series, animations, and the acclaimed film series.

3.4 Terminography

Terminography is a practical discipline closely linked to Terminology. Finatto (2014) examines the objective of this field of study as follows:

> Description of the linguistic, conceptual, and pragmatic properties of terminological units in one or more languages in order to produce reference works such as dictionaries, glossaries, vocabularies in print or electronic format, terminological databases, and specialized knowledge bases (Finatto, 2014, p. 439) .[4]

Therefore, Terminography can involve the production of technical vocabularies or glossaries for a specific field, organizing terms and their definitions systematically. It is an area that encompasses the search and systematization of relevant terms within a specific domain, so that other professionals can benefit from reference works that may be enriched by this effort.

3.5 Wordsmith Tools

Wordsmith Tools is a software developed by linguist Mike Scott in 1996 at the University of Liverpool. Even after so many years, it continues to be updated and, at the time of this writing, is in its version 9, released in January 2024. Published by Oxford University Press and Lexical Analysis Software, Wordsmith Tools integrates tools for textual analysis, allowing, for example, the study of word frequency, keyword listing, and identification of collocations. In its interface, Wordsmith Tools presents three main tools (as can be seen at the top of Figure 1): Concord, Keywords, and WordList.

[4] Descrição das propriedades linguísticas, conceituais e pragmáticas das unidades terminológicas de uma ou mais línguas, a fim de produzir obras de referência, tais como dicionários, glossários, vocabulários em formato papel ou eletrônico, bases de dados terminológicos e bases de conhecimento especializado."

Figure 1 - Wordsmith Tools 9.0

Source: WST, by Mike Scott (2024)

Berber Sardinha summarizes them well in the following paragraph:

> WordList: generates word lists containing all the words from the selected file or files, listed together with their absolute and percentage frequencies. It also compares lists, creating consistency lists that indicate how many lists each word appears in. Concord: performs concordances, or listings of a specific word (the 'node,' node word, or search word) along with the portion of text where it occurred. It also provides collocate lists, which are words that appear near the node word. Keywords: extracts words from a list whose frequencies are statistically different (higher or lower) from the frequencies of the same words in another corpus (reference corpus). It also calculates key keywords, which are key in multiple texts. (2009, p. 9)[5]

The software also provides secondary tools for corpus processing, such as the Text Converter and the Corpus Checker. According to Ribeiro (2004), the use of tools like Wordsmith Tools can still have certain limitations for translators due to the complexity of corpus compilation. Therefore, the existence of a resource like the Online Technical Vocabulary, or VoTec, by Fromm (2007) and its derivative projects (Fromm; Lisboa, 2024), proves to be important for this purpose.

[5] "WordList: produz listas de palavra contendo todas as palavras do arquivo ou arquivos selecionados, elencadas em conjunto com suas freqüências absolutas e percentuais. Também compara listas, criando listas de consistência, onde é informado em quantas listas cada palavra aparece. Concord: realiza concordâncias, ou listagens de uma palavra específica (o 'nódulo', node word ou search word) juntamente com parte do texto onde ocorreu. Oferece também listas de colocados, isto é, palavras que ocorrem perto do nódulo. KeyWords: extrai palavras de uma lista cujas freqüências são estatisticamente diferentes (maiores ou menores) do que as freqüências das mesmas palavras num outro corpus (de referência). Calcula também palavras-chave chave, que são chave em vários textos."

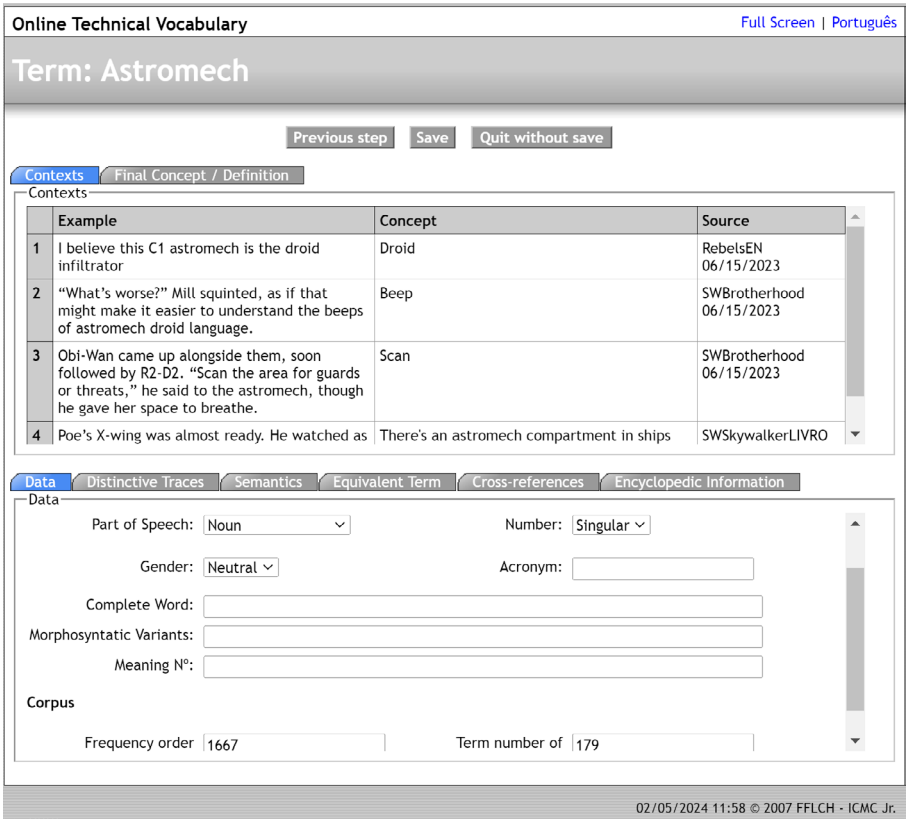
3.6 Terminology in Fiction

Terminology in Fiction, a derived version of the VoTec (Online Technical Vocabulary; available at: http://votec2.ileel.ufu.br) project (Fromm; Lisboa, 2024), emerges as a web-based terminological management environment aimed at assisting novice and professional translators in projects that demand heavy terminological translations (Fromm, 2007). The work is based on the premise that there is a significant need for integration between the fields of Translation and Terminology since, in the author's words, "increasingly, the translator also plays the role of terminologist and occasionally terminographer" (Fromm, 2007, p. 13).

The TF platform is a data bank supplied by researchers with terms previously gathered through corpora, processed by software such as Wordsmith Tools, for example. Along with the term candidates (enumerated from the WST's Keywords tool), the platform allows the researcher to input all contexts in which the word under analysis was used (from the WST's Concordance tool), enabling the registration of the term, the contexts that best explain its definition, and the final definition created by the researcher solely based on the selected contexts. The terms registered on the platform go through an approval process by the administrator, who must assess the quality of the registered contexts and their sources, distinctive features, encyclopedic information, the final concept, and the definition of each term.

For the completion of this work, the version of the TF environment shown in the figure 2 below was used.

Figure 2 - Astromech term in TF's databank (not published).



Source: TF databank

In general terms, the Corpus Linguistics approach used in this project is mainly based, as can be seen, on issues of Terminology (a word in a specific context of use, such as in a technical area), Terminography (development of products from terminological studies), the Communicative Theory of Terminology (interdisciplinarity and plurality in the treatment of what is considered a term), Ethnoterminology (the idea that any word can have the status of a term within a fictional universe), the lexical analysis software (WordSmith Tools) necessary to process the material from the chosen saga (Star Wars), and the platform (Terminology in Fiction) that we used to create bilingual entries. In the next section, we will analyze the processes of compilation, cleaning, systematization, and exploration of the corpus.

4 Methodology

The corpus for the project was compiled over the course of three months, with the English files being collected first, followed by the translated version of each one. In both languages, the corpus was divided into five parts: Animated Series/Animações, Books/Livros, Films/Filmes, TV Series/Séries, and Everything/Tudo, with the latter being the compilation of all the others – important for the creation of the WordList later.

In total, 760 files were downloaded, with 380 in the original language and 380 translated into Portuguese. For the corpus composition, only content considered canonical up to the time of compilation was selected, and it needed to be easily accessible. This excluded, for example, comic books, which despite being canonical, presented considerable difficulty in converting the issues into text documents.

Books constitute the most substantial part of the corpus, being nearly three times larger than the second-largest category, which is animations. This material corresponds to a total of 38 canonical novels that have been officially or unofficially translated into Portuguese. The selected works belong, within the Star Wars timeline, to the High Republic Era, the Republic Era, the Imperial Era, and the New Republic Era.

The entire Legends series, formerly known as the Expanded Universe, was disregarded, as were short stories, children's series, art and behind-the-scenes books, dictionaries, and encyclopedias. Clearly, books represent the richest portion of the Star Wars saga's world-building, with the first work being published even before the launch of the cinematic universe. Therefore, if all this data were considered for the study, the result would easily exceed a thousand files, significantly increasing the complexity of the compilation process.

In English, the books were found in online repositories. All of them were downloaded in EPUB format and later converted to TXT. This conversion presented some difficulties initially, as it couldn't preserve the character encoding of the original text correctly, which is a common issue when dealing with the diacritic accents of the Portuguese language.

Finding translations for the books into Portuguese was also a time-consuming process. The solution found to complete the research was to rely on the ongoing work of the non-professional group "Tradutores dos Whills," which translates all Star Wars literature not officially published in Brazil into Brazilian Portuguese. It was particularly interesting to analyze how the team maintained the original visual identity and graphic design, making it more accessible to read works that might never make it to Brazilian bookstores. Therefore, even though many works were officially translated and used in the corpus compilation, it is important to

mention the presence of non-professional translations in this project, which encompass not only books but all other categories as well.

The books that form the corpus are: A New Dawn, Aftermath, Aftermath: Empire's End, Aftermath: Life Debt, Ahsoka, Alphabet Squadron, Battlefront - Twilight Company, Battlefront II - Inferno Squad, Bloodline, Brotherhood, Catalyst, Dark Disciple, Heir to the Jedi, Leia, Princess of Alderaan, Lords of the Sith, Lost Stars, Master & Apprentice, Phasma, Queen's Peril, Queen's Shadow, Rebel Rising, Resistance Reborn, Rogue One, Tarkin, The Force Awakens, The High Republic - Into the Dark, The High Republic - Light of the Jedi, The High Republic - Midnight Horizon, The High Republic - Out of the Shadows, The High Republic - The Rising Storm, The High Republic – The Fallen Star, The Last Jedi, The Rise of Skywalker, Thrawn – Alliances, Thrawn – Treason, Thrawn Ascendancy - Chaos Rising, Thrawn Ascendancy - Greater Good, Thrawn Ascendancy - Lesser Evil.

After collecting the books, the process of searching for the constituent elements of the second most significant category in the corpus, which comprises animations, was initiated.

The animations that form the corpus are: one season of The Bad Batch, seven seasons of The Clone Wars, one season of Droids, two seasons of Ewoks, four seasons of Star Wars Rebels, two seasons of Star Wars Resistance, one season of Visions.

The expressiveness of this corpus arises from the longevity of some of these series, such as The Clone Wars, which reached a significant number of 133 episodes spread across seven seasons. Rebels, in turn, concluded after 75 episodes in four seasons. In terms of the number of files, the corpus is consequently the most substantial, with 304 subtitle files for each language.

In addition to animations, subtitles were also used as items for the composition of the film and TV series corpora.

Forming the film corpus are: Star Wars: Episode IV - A New Hope (1977), The Star Wars Holiday Special (1978), Star Wars: Episode V - The Empire Strikes Back (1980), Star Wars: Episode VI - Return of the Jedi (1983), Caravan of Courage: An Ewok Adventure (1984), Ewoks: The Battle for Endor (1985), Star Wars: Episode I - The Phantom Menace (1999), Star Wars: Episode II - Attack of the Clones (2002), Star Wars: Episode III - Revenge of the Sith (2005), Star Wars: The Clone Wars (2008), Star Wars: Episode VII - The Force Awakens (2015), Rogue One: A Star Wars Story (2016), Star Wars: Episode VIII - The Last Jedi (2017), Solo: A Star Wars Story (2018), Star Wars: Episode IX - The Rise of Skywalker (2019).

And the series: one season of The Book of Boba Fett, two seasons of The Mandalorian, one season of Obi-Wan Kenobi.

For live-action and animated series, all episodes released up to the time of corpus compilation were considered. The Mandalorian was in its second season, while Obi-Wan Kenobi and The Book of Boba Fett were in their first seasons. Likewise, regarding animations, The Bad Batch and Visions were in their first seasons. The Clone Wars was in its seventh season, Rebels in its fourth, and Resistance in its second. Ewoks and Droids were concluded in the 1980s. New releases from the Star Wars franchise for the Disney+ streaming platform, such as Ahsoka, Star Wars: Young Jedi Adventures, Tales of the Jedi, The Acolyte, and projects subsequent to these, were not included in the corpus as they had not yet been released during the compilation period.

4.1 Corpus Management

The corpora in both languages were divided into five parts. In Portuguese: "Desenhos", "Filmes", "Livros", "Séries", and "Tudo". In English: "Animated Series," "Films," "Books," "TV Series," and "Everything." All files from each corpus and their respective Wordlists were saved, separated by languages in distinct folders.

For corpus processing, the first step was to convert all subtitles and books to TXT format. For subtitles, SubtitleEdit was used, which is capable of converting and cleaning files in bulk. For books, online file conversion tools like Convertio were employed. A second conversion, this time for encoding, was necessary to use the corpus in WordSmith Tools. To achieve this, the software itself was used, specifically the Text Converter tool, which enables files to be converted into Unicode encoding.

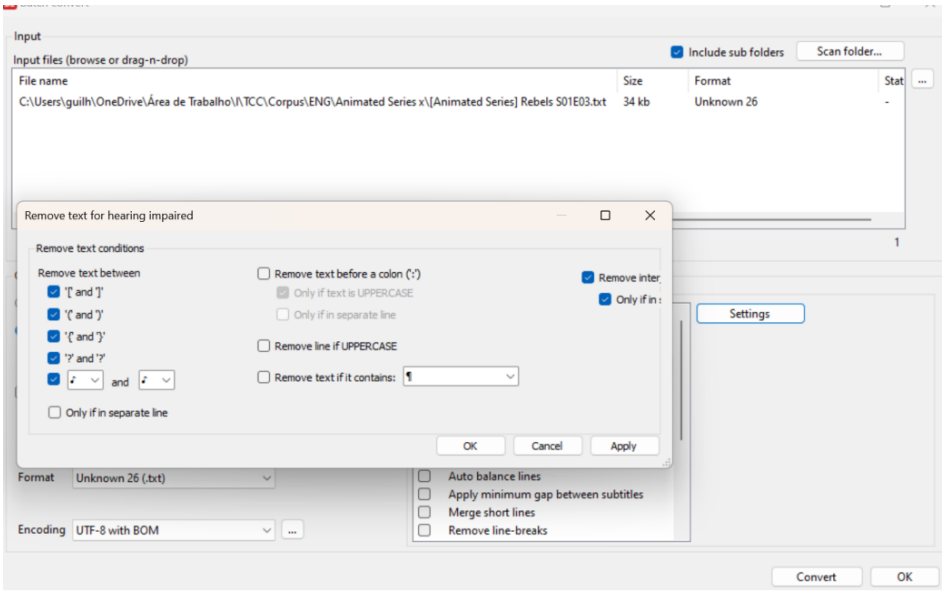
Failure in the corpus encoding can lead to processing errors, as illustrated below in Figure 3 (with what should be the word "didn't").

Figure 3 - Example of encoding error in WordSmith wordlist.

Source: extracted from Wordsmith Tools

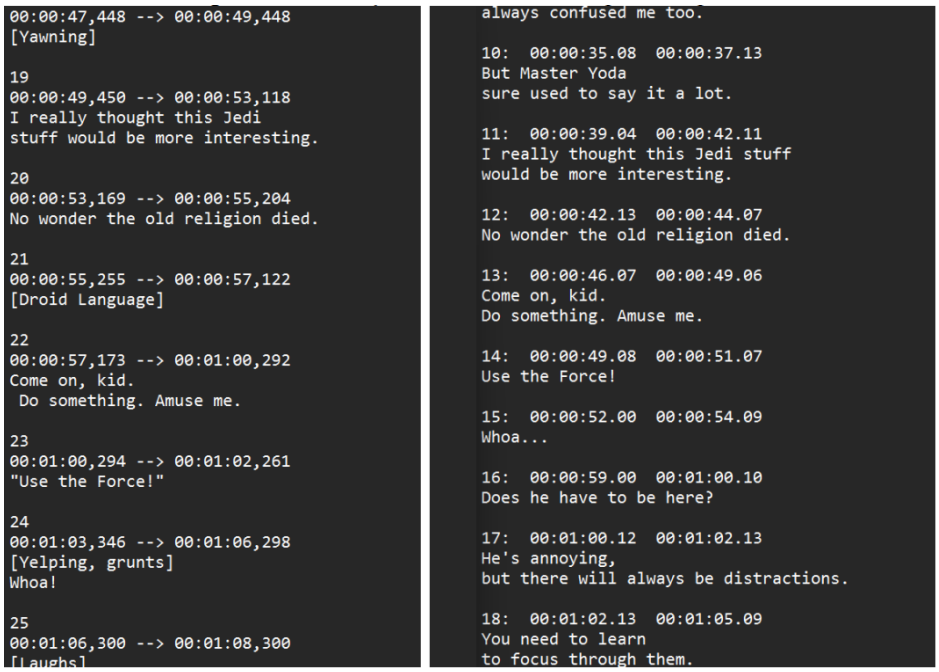
The corpus of animations, as well as series and films, consists of subtitles that were freely available on the internet. The English subtitles found are often typically labeled as "HI," meaning Hearing Impaired (for those with total or significant hearing loss). This means that these subtitles include audio descriptions that do not correspond to the dialogue, such as music, screams, cries, laughter, sounds, and different intonations. Since such descriptions are not the focus of our analyses and would therefore influence the data obtained, the English subtitles underwent an additional process to remove all this information. Initially, this was done manually, and later with the assistance of the SubtitleEdit software, which performs this process automatically (as seen in Figures 4 and 5).

Figure 4 - Subtitle Edit program configuration.



Source: Subtitle Edit

Figure 5 - Subtitle Edit results.

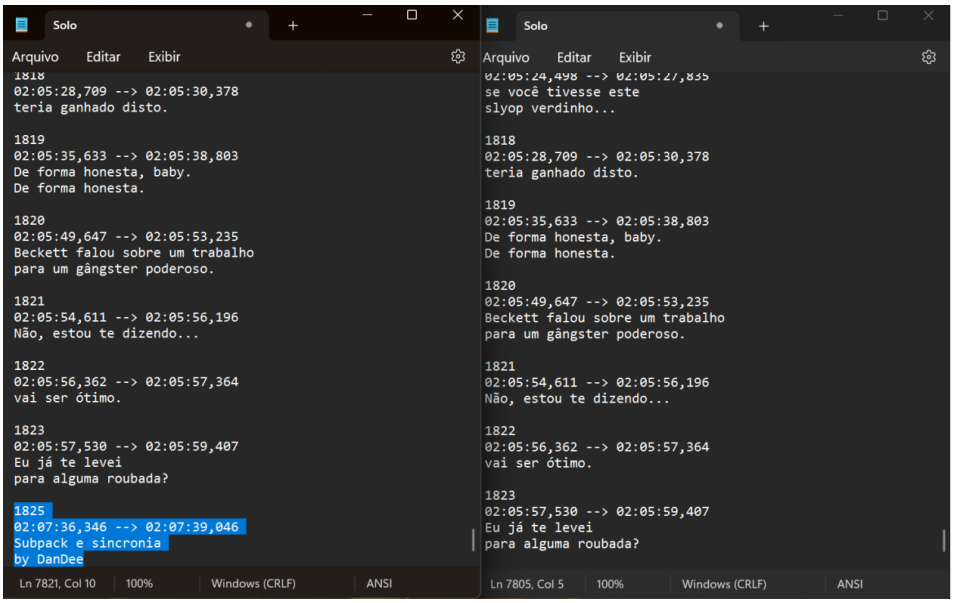


Source: screenshot by the author

In the case of translated subtitles, although HI elements are relatively rare, the items needed to undergo a cleaning process for a different reason. These subtitles, especially in the field of older animations and series, were produced by non-professional groups that credited their work in every episode two or three times. While understanding that the work of these teams was crucial for the completion of this research, the credits had to be removed to prevent the corpus from being polluted. Unlike HI elements, which are systematically enclosed

in square brackets within the subtitle, these credits do not follow a standard pattern, making it impossible for software like SubtitleEdit to perform automatic removal. Therefore, the cleaning of translated subtitles was a manual task (as shown below in Figure 6).

Figure 6 - Subtitle translator group advertised in the subtitle.



Source: screenshot by the author

4.2 Selection of term candidates

When conducting the corpus analysis with Wordsmith Tools, keywords are identified based on their frequency statistically compared to a reference corpus. These keywords stand out due to their significant occurrence in relation to the overall corpus, providing relevant insights into the key terms used in the analyzed texts – in this research, the canonical universe of Star Wars.

The terms chosen for registration in the TF platform were selected from the first 500 words in the KeyWords list generated by WordSmith Tools. To generate this list of keywords, a version of the COCA (Corpus of Contemporary American English) was used as a reference corpus, limited to its first 100,000 words. The result can be seen in Figure 7 below.

Figure 7 - first 33 keywords from the Star Wars corpus.

N	Key word	Freq.	%	Texts	RC. Freq.	%	BIC	Log_L	Log_R	P
1	BACK	9.445	0,20	360	6.366		64.181,05	64.200,92	7,07	0,0000000000
2	TIME	7.762	0,17	364	4.065		55.012,21	55.032,07	7,44	0,0000000000
3	DIDN	5.533	0,12	31	0		49.994,90	50.014,76	139,30	0,0000000000
4	THRAWN	5.392	0,12	22	0		48.720,35	48.740,21	139,26	0,0000000000
5	JEDI	5.215	0,11	242	0		47.120,38	47.140,25	139,21	0,0000000000
6	GOOD	4.646	0,10	362	113		40.910,84	40.930,70	11,87	0,0000000000
7	LL	4.171	0,09	31	0		37.683,29	37.703,16	138,89	0,0000000000
8	VE	3.843	0,08	33	0		34.718,39	34.738,25	138,77	0,0000000000
9	EYES	3.796	0,08	123	461		31.383,93	31.403,79	9,55	0,0000000000
10	LEIA	3.260	0,07	33	0		29.448,44	29.468,30	138,53	0,0000000000
11	DON	4.104	0,09	31	2.176		29.020,97	29.040,83	7,42	0,0000000000
12	WASN	3.189	0,07	31	0		28.806,65	28.826,51	138,50	0,0000000000
13	FIND	3.167	0,07	347	975		24.108,46	24.128,32	8,20	0,0000000000
14	CAPTAIN	2.625	0,06	178	18		23.493,35	23.513,21	13,69	0,0000000000
15	SHIP	5.839	0,13	285	26.683		22.729,52	22.749,38	4,31	0,0000000000
16	VOICE	3.371	0,07	71	2.580		22.363,94	22.383,80	6,89	0,0000000000
17	VADER	2.297	0,05	39	0		20.743,54	20.763,40	138,03	0,0000000000
18	SIDE	2.475	0,05	182	253		20.673,08	20.692,94	9,79	0,0000000000
19	DROID	2.452	0,05	201	256		20.455,55	20.475,41	9,76	0,0000000000
20	PADMÉ	2.255	0,05	22	0		20.363,89	20.383,75	138,00	0,0000000000
21	ANAKIN	2.237	0,05	115	0		20.201,18	20.221,04	137,99	0,0000000000
22	FELT	2.531	0,05	64	848		19.069,94	19.089,80	8,08	0,0000000000
23	THINGS	2.148	0,05	260	45		18.958,82	18.978,68	12,08	0,0000000000
24	COULDN	2.091	0,05	31	0		18.881,43	18.901,29	137,89	0,0000000000
25	SHIPS	2.237	0,05	142	397		17.976,46	17.996,33	9,00	0,0000000000
26	OBI	1.988	0,04	92	0		17.950,38	17.970,24	137,82	0,0000000000
27	ROOM	1.872	0,04	113	157		15.800,18	15.820,05	10,08	0,0000000000
28	MAN	1.927	0,04	160	288		15.693,40	15.713,27	9,25	0,0000000000
29	HADN	1.711	0,04	31	0		15.446,48	15.466,34	137,60	0,0000000000
30	DROIDS	1.735	0,04	178	78		15.021,75	15.041,62	10,98	0,0000000000
31	POE	1.583	0,03	24	0		14.289,44	14.309,30	137,49	0,0000000000
32	POWER	1.813	0,04	220	459		14.092,02	14.111,88	8,49	0,0000000000
33	TARKIN	1.506	0,03	22	0		13.593,41	13.613,27	137,42	0,0000000000

Source: KeyWords tool, WordSmith.

The terms extracted from the first 500 keywords were classified into nine specific semantic categories, which later became sub-areas in the TF database. The categories and pre-selected terms are indicated in table 1.

Table 1 – Gathered terms.

Vehicles	Species	Places	Groups	Weapons	Droids	Techno-logy	Profes-sions	Reli-gions
Starfigh-ter	Wookie	Coruscant	Rebels	Lightsaber	Astromech	Droid	Storm-trooper	Force
Star Destroyer	Ewok	Alderaan	Clones	Blaster	Protocol Droid	Hyperspace	Twilight Soldier	Sith
Spaceport	Hutts	Naboo	Separatists	Laser Cannons	Medical Droid	Hyperdrive	First Senator	Jedi
Imperial Star Destroyer		Manda-lore	Youngling		Server Droid	Datapad		
		Endor	Padawan		Pill Droid	HoloNet		
					Battle Droid	Lightspeed		
					Tactical Droid			
					Droid Tri Fighter			
					Moderator Droid			

Source: KeyWords tool from WordSmith.

In table one, we can see some of the terms found within the selected range of the first keywords for each category. For subsequent registration in TF, two representative examples from each category were chosen to best illustrate the use of the platform and the usefulness of the proposal.

4.3 Exploration of concordance lines

The exploration of terms was made possible through the Concord tool in WordSmith Tools, which allows researchers to identify the contexts in which a specific word or expression occurs in the corpus. For example, if a researcher is working with a Star Wars corpus in English, they can find all the contexts in which the word "Force" was used, whether they are cultural, scientific, religious, etc. Furthermore, it is also possible to analyze the words that surround the term, its cotext, which can even lead to the discovery of new technical phraseology[6] – as was the case with "Imperial Star Destroyer," which demonstrated relevance during the exploration of Star Destroyer contexts.

For registration in TF's databank, it is necessary to gather contextual examples that demonstrate different concepts related to the studied term, allowing the creation of a terminological definition based on the word in use.

In figure 8 we can find examples of contexts to define the term "Ewok":

Figure 8 - Concordance lines for the term EWOK in Portuguese.

N	Concordance	Set	Tag	Word #	Sent #	Sent Pos	Para #	Para Pos	H #	H Pos	Sect #	Sect Pos	File	Date	%
49	Ewok. 99: 00:07:43.04 00:07:45.15 Quer dizer que **o Ewok que** vimos... 100: 00:07:45.17 00:07:49.03			817	94	9	0	816			0	816	[Desenho] Ewoks S01E13.txt	2022/nov/14 00:	35%
50	. 242: 00:18:21.06 00:18:23.22 Esse não pode ser **o Ewok que** me derrotará. 243: 00:18:26.22 00:18:			1.821	252	10	01	0			01	0	[Desenho] Ewoks S02E06.txt	2022/nov/14 00:	82%
51	gorjeta. 319: 00:22:20.13 00:22:24.11 Acho que **os Ewok fazem** mal para nossos negócios. 320: 00:22:			2.564	324	8	02	3			02	3	[Desenho] Ewoks S01E05.txt	2022/nov/14 00:	95%
52	02:48.20 00:02:52.09 Pronto. Kneesaa, você é **outra Ewok.** 41: 00:02:52.11 00:02:54.14 Sim. Mas qual?			368	30	6	0	367			0	367	[Desenho] Ewoks S02E13.txt	2022/nov/14 00:	15%
53	Chirpa. 43: 00:04:13.08 00:04:15.18 Nem **pensar, Ewok.** 44: 00:04:16.11 00:04:19.19 Venha e se junte			353	33	7	0	352			0	352	[Desenho] Ewoks S02E04.txt	2022/nov/14 00:	15%
54	:26.05 00:17:28.07 O que faremos com **a pequena Ewok? 216:** 00:17:28.20 00:17:30.00 É a Kneesaa!			1.669	221	11	01	8			01	8	[Desenho] Ewoks S02E08.txt	2022/nov/14 00:	74%
55	: 00:08:41.09 00:08:46.16 Então, minha **pequena Ewok, não** é divertido ser um Jinda? 137: 00:08:			1.146	116	8	01	5			01	5	[Desenho] Ewoks S01E05.txt	2022/nov/14 00:	42%
56	– uma pequena bolota dada a ela pelo **pequeno Ewok conhecido** como Wicket. Ela cultivou a planta			60.270	5...3	19	03	9			03	9	[Livro] Aftermath - Divida de Honra.txt	2022/out/14 00:	91%
57	47: 00:04:33.13 00:04:37.07 Agora, **pequeno Ewok, junte-se** ao meu totem 48: 00:04:39.19 00:04:			379	37	7	0	378			0	378	[Desenho] Ewoks S02E04.txt	2022/nov/14 00:	17%
58	. 69: 00:05:51.18 00:05:56.07 - Teebo? - **Pequeno Ewok, junte-se** ao meu totem. 70: 00:05:57.20 00:05			553	65	4	0	552			0	552	[Desenho] Ewoks S02E04.txt	2022/nov/14 00:	24%
59	. 134: 00:10:07.19 00:10:12.08 Nada mau, **pequeno Ewok, mas** tenho mais azar para você. 135: 00:10:			1.139	139	8	01	8			01	8	[Desenho] Ewoks S02E11.txt	2022/nov/14 00:	44%
60	a árvore-santuário que lhe fora dada pelo **pequeno Ewok Wicket.** Ela nunca foi capaz de sentir a árvore			77.817	7...6	14	07	6			07	6	[Livro] Aftermath - Fim do Império.txt	2022/out/14 00:	117%
61	:14:06.05 Eles devem ter lhe dado aquele **pirralho Ewok 174:** 00:14:06.07 00:14:08.03 para acalmá-la.			1.439	166	12	01	8			01	8	[Desenho] Ewoks S01E12.txt	2022/nov/14 00:	64%
62	217: 00:17:57.08 00:18:02.03 Que linda **princesa Ewok. 218:** 00:18:02.05 00:18:03.10 Não. 219: 00:			1.974	263	8	01	3			01	3	[Desenho] Ewoks S02E02.txt	2022/nov/14 00:	79%
63	214: 00:17:43.13 00:17:46.13 Temos uma **princesa Ewok para** a Rainha Lesmona. 215: 00:17:47.23 00:			1.955	260	8	01	4			01	4	[Desenho] Ewoks S02E02.txt	2022/nov/14 00:	78%
64	200: 00:16:37.01 00:16:40.21 Nossa, uma **princesa Ewok para** a Rainha Lesmona. 201: 00:16:45.04 00:			1.836	243	8	01	5			01	5	[Desenho] Ewoks S02E02.txt	2022/nov/14 00:	73%
65	:25.03 00:13:30.00 - Vão, tragam-me uma **princesa Ewok. - Sim,** Rainha Lesmona. 163: 00:13:35.16 00:			1.504	198	10	01	3			01	3	[Desenho] Ewoks S02E02.txt	2022/nov/14 00:	60%
66	, 101: 00:06:58.09 00:07:01.06 uma figura de **proa Ewok. 102:** 00:07:01.13 00:07:04.22 - Ewok? - Dá			762	93	14	0	761			0	761	[Desenho] Ewoks S02E09.txt	2022/nov/14 00:	34%
67	. 202: 00:13:49.04 00:13:51.12 Levaria se **pudesse, Ewok. 203:** 00:13:51.14 00:13:53.09 Mas,			1.709	194	8	01	8			01	8	[Desenho] Ewoks S01E05.txt	2022/nov/14 00:	62%
68	. 214: 00:14:02.20 00:14:04.15 Você disse **qualquer Ewok. 215:** 00:14:05.02 00:14:07.01 Tudo bem, se			1.692	231	8	01	1			01	1	[Desenho] Ewoks S02E07.txt	2022/nov/14 00:	66%
69	:58.02 00:13:01.23 Aposto que transformo **qualquer Ewok em** um guerreiro em um dia. 197: 00:13:02.13			1.555	211	9	01	4			01	4	[Desenho] Ewoks S02E07.txt	2022/nov/14 00:	61%
70	: 00:10:54.15 00:10:56.09 - Sabão Ewok! - **Sabão Ewok! 150:** 00:10:56.19 00:10:59.10 Vamos, rápido.			1.218	145	4	01	7			01	7	[Desenho] Ewoks S01E02.txt	2022/nov/14 00:	53%
71	Peguem! 252: 00:19:16.06 00:19:18.05 O **sabão, Ewok! 253:** 00:19:20.17 00:19:22.17 Entregue,			2.046	258	7	02	5			02	5	[Desenho] Ewoks S01E02.txt	2022/nov/14 00:	89%
72	se limpando. 95: 00:06:47.02 00:06:48.09 **Sabão Ewok. 96:** 00:06:48.22 00:06:50.19 Não há nada			800	86	6	0	799			0	799	[Desenho] Ewoks S01E02.txt	2022/nov/14 00:	35%
73	temos aqui? 149: 00:10:54.15 00:10:56.09 - **Sabão Ewok! - Sabão Ewok!** 150: 00:10:56.19 00:10:59.10			1.215	144	7	01	4			01	4	[Desenho] Ewoks S01E02.txt	2022/nov/14 00:	53%
74	:51:28.08 O cristal pertence a Kaink, **a sacerdotisa Ewok. 361:** 00:51:31.18 00:51:34.12 Mas antes que			2.805	462	12	02	4			02	4	[Filme] Caravana da Coragem.txt	2022/nov/14 00:	65%
75	59: 00:05:12.05 00:05:17.11 Um monstro **selvagem Ewok com** seis braços 60: 00:05:17.18 00:05:21.02			481	51	8	0	480			0	480	[Desenho] Ewoks S01E13.txt	2022/nov/14 00:	21%
76	frito. 45: 00:03:25.07 00:03:26.17 Eu **também, Ewok. 46:** 00:03:31.02 00:03:34.00 Mas alguns			370	33	7	0	369			0	369	[Desenho] Ewoks S01E08.txt	2022/nov/14 00:	17%
77	seja. Arsad sorri. – Eu poderia colocá-lo na **terapia Ewok, em** vez disso. Algumas das criaturas nativas			99.383	9...2	8	09	2			09	2	[Livro] Aftermath - Divida de Honra.txt	2022/out/14 00:	150%
78	:35.09 00:16:38.07 É melhor fazer um bom **trabalho, Ewok. 247:** 00:16:38.09 00:16:40.01 Senão. 248: 00:			2.049	240	11	02	8			02	8	[Desenho] Ewoks S01E05.txt	2022/nov/14 00:	75%
79	se reúnem para uma cerimônia **tradicional Ewok. 301:** 00:42:19.05 00:42:22.22 Antes de			2.330	398	13	02	9			02	9	[Filme] Caravana da Coragem.txt	2022/nov/14 00:	54%
80	: 00:12:06.21 00:12:09.06 É melhor olhar para **trás, Ewok. 191:** 00:12:17.16 00:12:20.16 Kvark. Corram.			1.591	211	10	01	0			01	0	[Desenho] Ewoks S01E07.txt	2022/nov/14 00:	57%
81	27: 00:01:59.21 00:02:04.03 Saudações da **tribo Ewok, Rainha** das Wisties. 28: 00:02:16.15 00:02:			244	13	8	0	243			0	243	[Desenho] Ewoks S01E01.txt	2022/nov/14 00:	11%
82	: 00:18:18.16 00:18:23.01 Lutaremos até o **último ewok para** proteger nossas árvores da alma. 275: 00			2.177	262	9	02	6			02	6	[Desenho] Ewoks S01E10.txt	2022/nov/14 00:	84%

Source: Concordance, WordSmith Tools

For this term, five contexts were selected, each one supplying different concepts. The sources vary; in the case of "Ewok," the excerpts are from the movie "Caravan of Courage: An Ewok Adventure" from 1984 and the animated series "Ewoks" from 1985. From the chosen con-

[6] "There are various notions of what a phraseologism is, but in this text, I consider it to be any grouping of two or more words whose combined frequency is highlighted by lexical analysis programs." (Fromm, 2020) Original: Existem várias noções do que seja um fraseologismo, mas considero o mesmo, neste texto, como qualquer agrupamento de duas ou mais palavras cuja frequência combinada seja destacada pelos programas de análise lexical."

texts, it was possible to discover that Ewoks are a tribe of friends living in trees in the forests of Endor, with their own ceremonies, warriors, and wizards.

In the figure 9, we can analyze the selected contexts in order to create a definition.

Figure 9 - Selected contexts for the term EWOK.

Registered contexts

Example	Concept	Source	Actions
"Gather for a traditional Ewok ceremony."	Ceremonies	CaravanEwok 04/13/2023	edit - delete
"Ewoks! We're the Ewoks, raggedy Ewoks Living in the tall trees Living in the spiral Dancing in the forest On the moon of Endor Ewoks all together And we're having fun Friends together Friends forever Ewoks We're careless, little Ewoks We like adventure Helping friends in danger Out in the forest Sharing in the magic On the moon of Endor Ewoks all together And we're having fun Friends together Friends forever Ewoks We're the Ewoks, yeah"	Friends living in trees in a forest of Endor	Ewoks 04/13/2023	edit - delete
"It's Queen Izrina!" "Greetings from the Ewok tribe, Queen of all the Wisties."	Tribe	Ewoks 04/13/2023	edit - delete
"And the Ewok warriors will stop you good if you even show your face near there."	Warriors	Ewoks 04/13/2023	edit - delete
"But, Latara, no Ewok wizard can do that. Not since the Crystal Cloak was stolen."	Wizards	Ewoks 04/13/2023	edit - delete

registered contexts: 5

Source: TF databank

The use of Concord is not only useful for the elaboration of term definitions but also, among various uses, for exploring translation aspects – one of the themes of this work. Let's return to the example of "Star Destroyer". 385 different contexts of "Star Destroyer" were found in the English corpus, compared to only 21 in the Portuguese corpus. The clear discrepancy demonstrates that there were issues about its translation, and it was up to the researcher to verify one of the English contexts and locate the same passage in the Portuguese version to find the translated term. As a result, it was discovered that the term historically was more frequently translated as "Destróier Estelar" in Brazilian materials. Furthermore, 21 contexts were found where a term was either not mentioned in Brazilian materials or simply not translated, emphasizing the need for a terminological research source like TF, since the original title, in a foreign language, was retained - in contrast to the established standard by the Star Wars Brazilian translation.

After all these steps, the terms were ready for registration in TF. In the next section, we analyze how this stage was carried out.

5 The development of Star Wars entries in the vocabulary

As previously explained, the general objective of the project is to create a proposal for a Star Wars vocabulary in the TF platform. The query page of the TF project can be accessed for free from any computer or mobile device, making it a simplified environment for translators to consult.

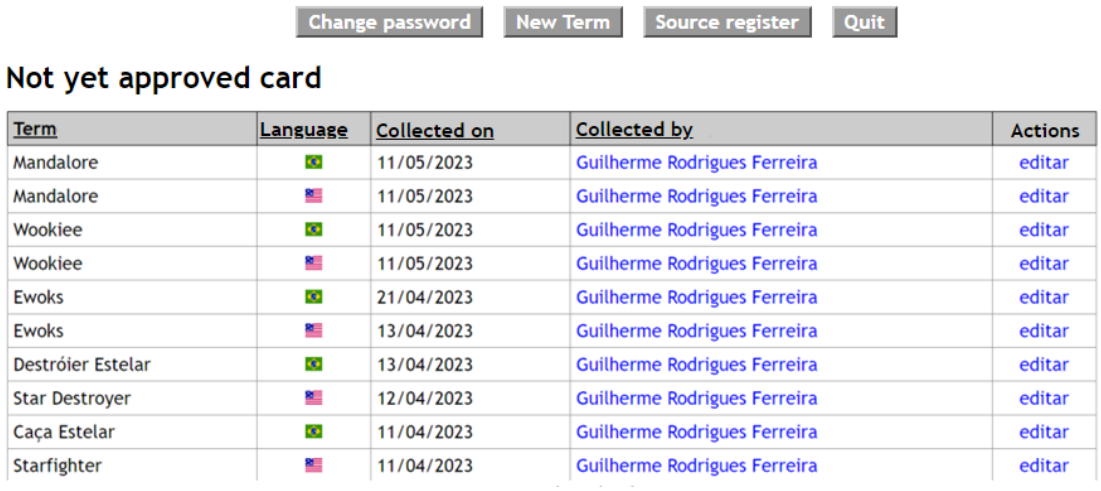
5.1 Creation of the Star Wars project in TF

The project was created within the field of Science and Fiction Terminology, where there are already studies on works like "Farscape", "Doctor Who", and "Star Trek", among others. However, this project was the first to include sub-areas, which further refines the research process. This means that when searching for terms, users can navigate through the categories as shown in Table 1.

5.2 Building the entries

Researchers who are TF users have access to the data bank page where they can register new terms and contribute to the platform. On the home page, as you can see in figure 10, researchers can access a list of all terms that have been registered but not yet approved or awaiting review. They also have the options to create a term and register a new source.

Figure 10 - Creating a term in TF.

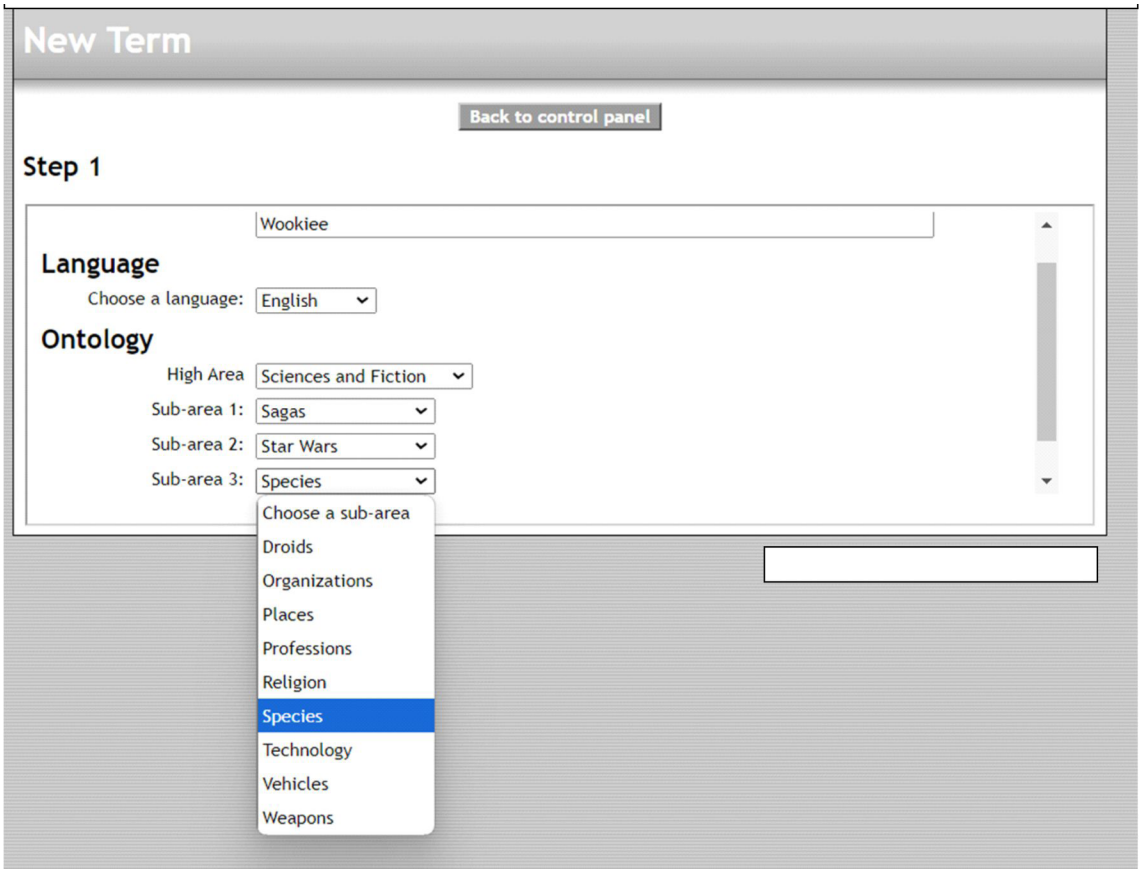


Term	Language	Collected on	Collected by	Actions
Mandalore		11/05/2023	Guilherme Rodrigues Ferreira	editar
Mandalore		11/05/2023	Guilherme Rodrigues Ferreira	editar
Wookiee		11/05/2023	Guilherme Rodrigues Ferreira	editar
Wookiee		11/05/2023	Guilherme Rodrigues Ferreira	editar
Ewoks		21/04/2023	Guilherme Rodrigues Ferreira	editar
Ewoks		13/04/2023	Guilherme Rodrigues Ferreira	editar
Destróier Estelar		13/04/2023	Guilherme Rodrigues Ferreira	editar
Star Destroyer		12/04/2023	Guilherme Rodrigues Ferreira	editar
Caça Estelar		11/04/2023	Guilherme Rodrigues Ferreira	editar
Starfighter		11/04/2023	Guilherme Rodrigues Ferreira	editar

Source: TF databank

The first step in creating a term is to name it. Some words have the possibility of being in singular or plural form, as is the case with "Ewoks" in figure 10, which was not registered as "Ewok" in the singular form. For this work, we considered the most frequent form in the corpus for registration in TF – which means that in the entire canonical universe of Star Wars, the usage of the term "Ewoks" is more frequent than "Ewok." However, both forms (singular and plural) were subsequently lemmatized, meaning they were grouped in the corpus as a single term for analysis. This explains the possibility of finding contextual examples in the singular form of a term that was registered in the plural. After creating the term, it is necessary to choose the language, the major area, subarea 1, and subarea 2. The first belongs to the material being worked on, in our case, Star Wars. The second consists of the categories that were previously registered. Figure 11 shows how this process is done, starting with the registration of the term "Wookiee", for example.

Figure 11 - Creating the term WOOKIE

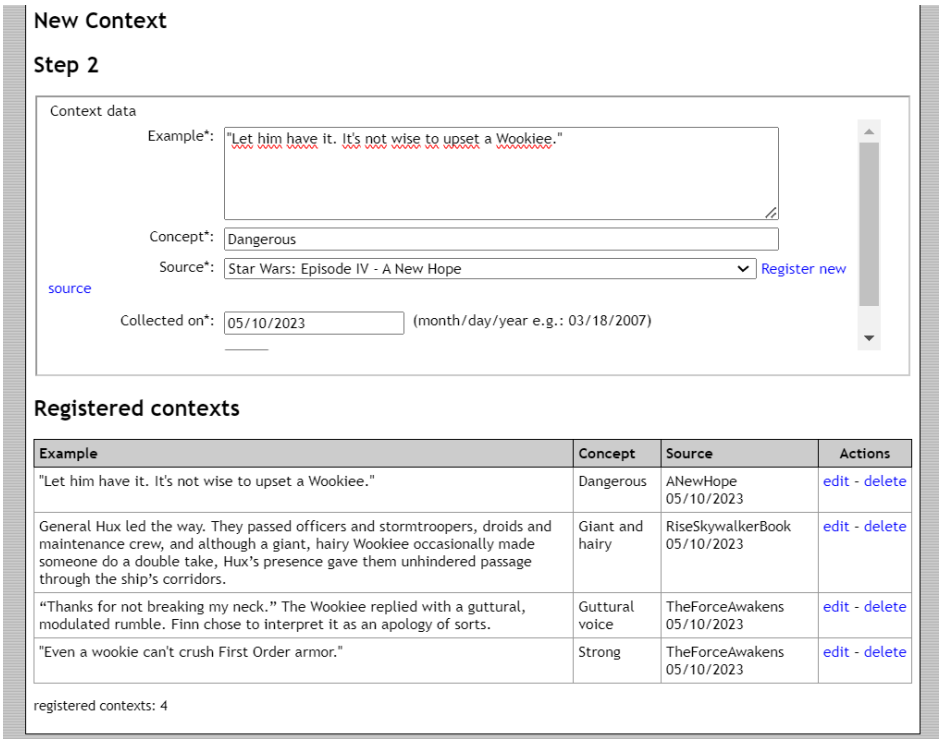


Source: TF databank.

The second step is to register the contexts. The definition of the term will be created solely from data extracted from real contextual fragments, as preconized by the VoTec general project (Fromm, 2007). As previously explained, with the help of the Concord tool, the researcher can search for a specific term in the corpus and analyze all the situations in which it was used. The goal is to capture a variety of concepts.

The context registration screen in TF can be seen in figure 12.

Figure 12 - Registration of the examples and contexts for the WOOKIE term



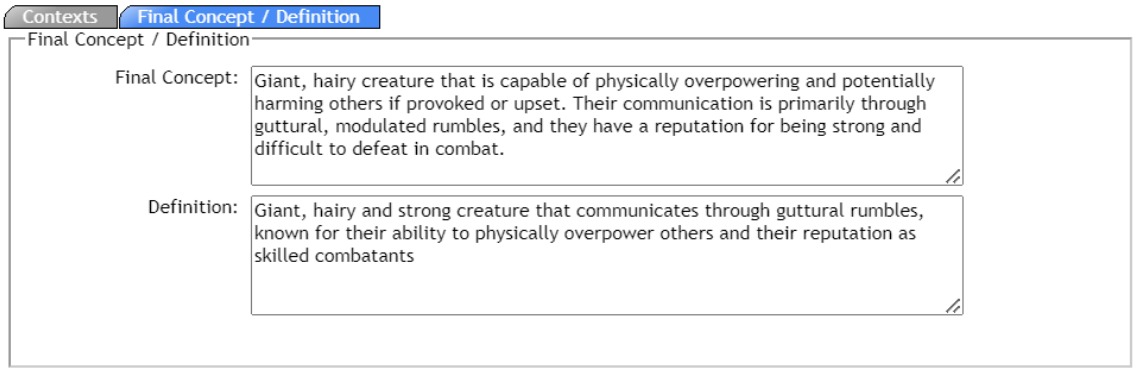
Source: TF databank.

As soon as a passage in the corpus that adequately explains the term is selected, that excerpt is inserted into the "Example" field. The second context registered in Figure 12, for example, serves to showcase the physical characteristics of a Wookiee – this is a kind of concept, and the more concepts we have, the richer the definition will be. Additionally, the source and collection data are also entered.

The next step involves providing details about the term and creating a concept and definition. We enter the grammatical category, number (singular or plural), gender (neutral in English, masculine/feminine in Portuguese), position in the frequency order, and the number of occurrences of the term. We also include the equivalent term in another language, if it has already been registered in TF, along with encyclopedic information.

Based on the selected excerpts, we create a final concept that effectively explains what the term is. The definition, in turn, must provide a single sentence that offers a polished and concise explanation developed from the final concept. The difference between the final concept and the definition can be seen in the figure 13, within the same example.

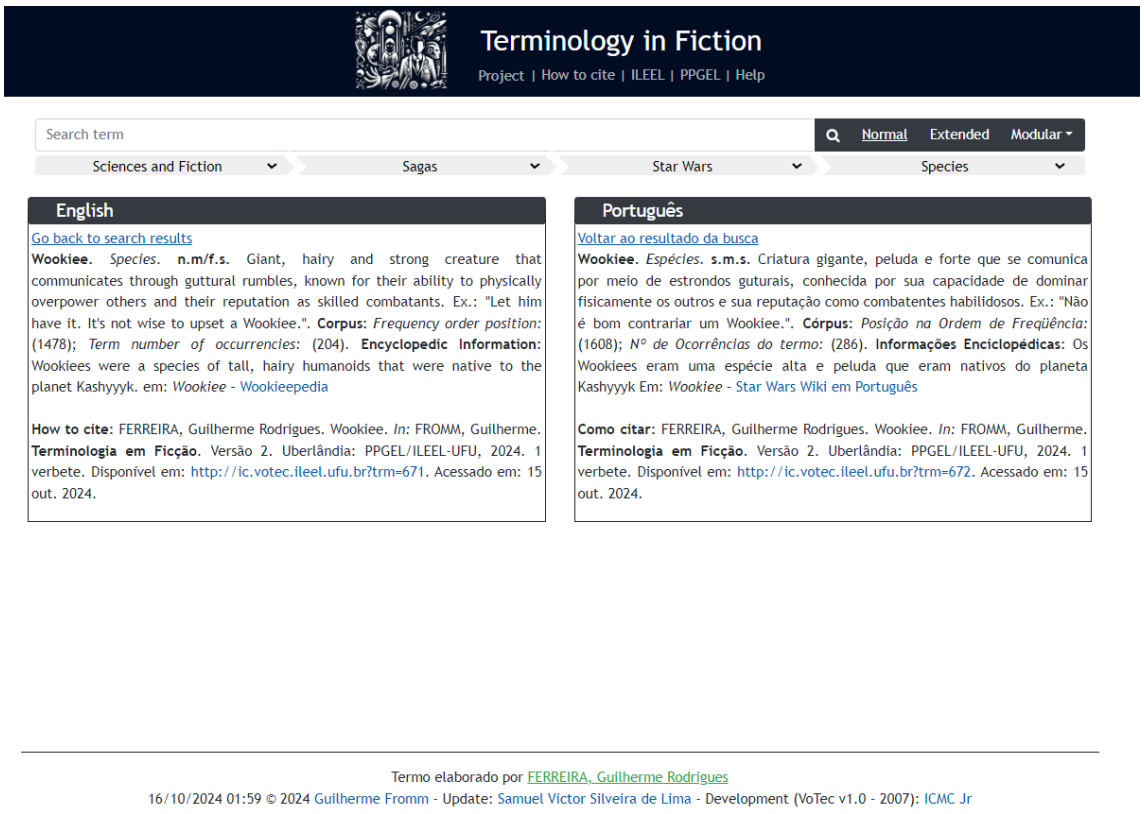
Figure 13 - Creating the final concept and the definition of the term WOOKIEE



Source: FT databank.

With this entire process completed, the term is ready to be submitted to the administrator, who must then evaluate and approve (or not) the work. For this project, this process was carried out 36 times, as two terms were selected and registered twice (in Portuguese and English) for each of the nine categories. In Figure 14, you can see the interface of TF for the end user.

Figure 14 - Term WOOKIEE, as shown in TF



Source: TF query page.

The final result can be found on the TF query page (http://ic.votec.ileel.ufu.br), where all approved terms are made available for consultation by the community. To the end user, as

shown in Figure 14, it is their role to input the areas they would like to explore, search for a term, and analyze the results in the provided languages (Portuguese and English). It is important to notice that the arrangement of the terms shown in Figure 14 are monolingual in contrast.

6 Final Remarks

The entire project presented here is based on our shared assumption (within the Ethnoterminology theory) that many fictional universes can contain a large number of terms, created or re-created (from words already existing in any language) for the narrative purposes of authors in their works (as demonstrated by Oliveira, 2016, in relation to the Harry Potter books). Especially in long sagas (literary or audiovisual ones), the use or creation of terms (even if they do not exist in the real world) is important to differentiate that story from others already told and to maintain verisimilitude in the work as a whole. In an era of mass streaming and a flood of audiovisual products, documenting how terminological elaboration is created in large sagas, such as Star Wars, and the consequent development of a bilingual reference work, can help translators and screenwriters in their work and also serve as a more elaborate database for fans of these sagas.

The studies from Corpus Linguistics have provided a more facilitated analysis of linguistic patterns through the examination of corpora, which can result in the development of tools and applications or even sociolinguistic studies. The relationship between Corpus Linguistics and Translation is longstanding, as highlighted by Kruger, Wallmach, and Munday (2011), and the ability to compare corpora in different languages has become commonplace for researchers and translators.. Building upon this principle, two corpora from the Star Wars canonical universe in Portuguese and English were compiled, with the aim of assisting translators and translation students with their works or studies. We also considered the cultural impact and relevance of the franchise, which not only involves the translation of its elements in various releases each year but also in all other media that in some way reference or will reference this material, such as Star Wars episodes in Family Guy, The Simpsons, and South Park.

As a prototype of a bilingual English/Portuguese Star Wars corpus-based vocabulary, the research objectives were fulfilled since the proposal is available on TF query page in a detailed manner, enabling translators who are dealing with the Star Wars universe in their work and research to consult it. However, this studycan be further extended by continuing the corpus analysis and its entries in TF. To work with the entire Star Wars terminology, more time would be required to enter all the terms and update the corpora, as the franchise is active and new content is continuously being released. This project also excluded canonical media that are difficult to transpose into corpora, such as comics and video games, which would be relevant in a more in-depth study of this same material.

In conclusion, the Star Wars franchise offers a considerably rich universe for corpus exploration, allowing analyses in the numerous fields of knowledge as literature, technology, etc., which have been created since the late 1970s. In this paper, we followed a more technical path, focusing on the construction of a prototype vocabulary with the intention of assisting the professional and amateur activities of translators who can benefit from TF.

Authorship declaration

Both authors worked on the writing of this article. The first author developed the original studies on his TCC under the supervision of the second author and using his platform. The revision of the article was also performed by both authors.

References

BARBOSA, M. A. Para uma etno-terminologia: recortes epistemológicos. *Ciência & Cultura*, v. 58, n. abr./maio/ju 2006, p. 48-51, 2006.

CABRÉ, M. T. *La Terminología*: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada, 1999.

CARNEIRO, R. M. O. *Discurso literário de fantasia infantojuvenil*: proposta de descrição terminológica direcionada por corpus. 2016. 281 f. Dissertação (Mestrado em Estudos Linguísticos) - Programa de Pós-graduação em Estudos Linguísticos, Universidade Federal de Uberlândia, Uberlândia, 2016. DOI http://doi.org/10.14393/ufu.di.2016.445

DIAS, C. A. Terminologia: conceitos e aplicações. *Ciência da Informação*, Brasília, v. 29, n. 1, p. 90-92, abr. 2000. IBICT. DOI http://dx.doi.org/10.1590/s0100- 19652000000100009.

FERREIRA, G. *Protótipo de um vocabulário bilíngue inglês/português de Star Wars baseado em Corpus*. 2022. 39 f. Trabalho de Conclusão de Curso (Graduação em Letras – Inglês e Literaturas de Língua Inglesa) – Universidade Federal de Uberlândia, Uberlândia, 2023. Disponível em: https://repositorio.ufu.br/handle/123456789/38772. Acesso em: 09 set. 2024.

FINATTO, M. J. B. Orientações para a terminografia: das teorias às práticas em busca de amplitude da informação terminológica. *In*: ISQUERDO, A. N.; CORNO, G. M. dal. (org.). *As Ciências do Léxico*: lexicologia, lexicografia e terminologia, Campo Grande. MS: Ed.UFMS, vol. VII, 2014, p. 439-459.

FROMM, G. *Terminologia em Ficção*. Versão 2. Uberlândia: PPGEL/ILEEL-UFU, 2024. 1 plataforma. Disponível em: http://ic.votec.ileel.ufu.br/. Acesso em: 09 set. 2024.

FROMM, G.; LISBOA, J. VoTec terminographic environment over the years: brief overview. *Acta Scientiarum. Language and Culture*, v. 45, n. 2, p. e67669, 23 fev. 2024. DOI: https://doi.org/10.4025/actascilangcult.v45i2.67669

FROMM, G. *VoTec*: a construção de vocabulários técnicos eletrônicos para aprendizes de tradução. 2007. 215f. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês). Departamento de Letras Modernas, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007. DOI: https://doi.org/10.11606/T.8.2008.tde-08072008-150855

KRUGER, A.; WALLMACH, K.; MUNDAY, J. *Corpus-Based Translation Studies*: research and applications. Londres: Continuum, 2011.

LARA, M, L, G, de. Linguagem documentária e terminologia. *Transinformação*, Campinas, v. 16, n. 3, p. 231-240, dez. 2004. DOI http://dx.doi.org/10.1590/s0103-37862004000300003.

LARA, M. L. G. de. *Elementos da Terminologia*: (apostila para uso didático). São Paulo, 2005.

LATORRE, V. R. D.. A dialética entre os extremos: da terminologia à etnoterminologia. *Caderno Seminal*, [S.L.], v. 19, n. 19, p. 70-94, 29 ago. 2013. DOI http://dx.doi.org/10.12957/cadsem.2013.12062.

PEIXOTO, L. M. Identificação de unidades fraseológicas no vocabulário de Star Trek: abordagens corpus-driven e corpus-based. *Domínios de Lingu@gem*, Uberlândia, v. 8, n. 2, p. 139–163, 2014. DOI: 10.14393/DL16-v8n2a2014-8.

RIBEIRO, G. C. B. Tradução técnica, terminologia e lingüística de corpus: ferramenta Wordsmith Tools. *Cadernos de Tradução*, Rio de Janeiro, v. 2, n. 14, p. 160-174, jan. 2004.

SARDINHA, T. B. *Pesquisa em Lingüística de Corpus com Wordsmith Tools.* Campinas: Mercado de Letras, 2009.

SCOTT, M. *WordSmith Tools version 9 (64 bit version*) Stroud: Lexical Analysis Software, 2024.

Star Wars: Episódio IV – Uma Nova Esperança (Comentários em áudio de George Lucas, Carrie Fisher, Ben Burtt e Dennis Muren). Direção de George Lucas. Estados Unidos: Twentieth Century-Fox Film Corporation, 2011. Blu-ray (121 min)

ZAMORA, R. E. M. Sobre tradução e terminologia das ciências sociais e humanas: quando a cultura encontra a "cultura". *Mutatis Mutandis*: Revista Latinoamericana de Traducción, [S. L.], v. 8, n. 2, p. 548-566, 2015.