UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE LETRAS
DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO
LÍNGUAS ESTRANGEIRAS APLICADAS AO MULTILINGUISMO E À SOCIEDADE DA INFORMAÇÃO

JEISIANE DA SILVA OLIVEIRA

EXTRAÇÃO TERMINOLÓGICA MULTILÍNGUE DA ÁREA DA MODA

Brasília/DF
2023

Jeisiane da Silva Oliveira

EXTRAÇÃO TERMINOLÓGICA MULTILÍNGUE DA ÁREA DA MODA

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília, como requisito para a obtenção do título de bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Orientador: Prof. Dr. Thiago Blanch Pires

Brasília/DF
2023

Jeisiane da Silva Oliveira

EXTRAÇÃO TERMINOLÓGICA MULTILÍNGUE DA ÁREA DA MODA

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília, como requisito para a obtenção do título de bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação.

Orientador: Prof. Dr. Thiago Blanch Pires

BANCA EXAMINADORA

_____

Prof. Dr. Thiago Blanch Pires (Orientador)

_____

Profa. Dra. Clarissa Prado Marini

_____

Profa. Dra. María del Mar Paramos Cebey

Resumo: Este trabalho tem como objetivo extrair e analisar termos multilíngues da área da moda nos idiomas português, inglês, espanhol e francês. O estudo justifica-se devido à importância desta indústria no Brasil, pois o país é um importante produtor têxtil e possui a maior cadeia têxtil do Ocidente, que vai desde a plantação de algodão e a confecção de peças, até o varejo e os desfiles de moda (ABIT - ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL E DE CONFECÇÃO, 2023). Os termos foram extraídos usando um corpus comparável de língua oral construído durante o percurso deste trabalho a partir da transcrição automática de 320 vídeos no YouTube, sendo 80 vídeos para cada idioma. Após extraídos, os termos foram analisados quanto à frequência no corpus, pertinência na área da moda, contextos em que aparecem, definições e os colocados que os acompanham. Com essas informações coletadas, foi possível identificar as equivalências entre os idiomas, partindo de seus contextos e definições (GODOY, 2019). Todas essas informações foram organizadas em um dossiê terminológico (PAVEL; NOLET, 2002) e posteriormente passadas às fichas terminológicas (KRIEGER, 2004). Este trabalho apresenta um recorte dos 30 termos com maior frequência no corpus e seus 74 colocados, ao finalizar a identificação de equivalências, encontrou-se 10 termos que apresentaram equivalências nos 4 idiomas, 20 que apresentaram equivalências em até 3 idiomas, 26 que apresentaram equivalências em até 2 idiomas e 45 que não apresentaram equivalências.
Palavras-chave: Terminologia, Linguística de Corpus, Moda, Multilinguismo.

Abstract: *This paper aims to extract and analyze multilingual fashion terms in Portuguese, English, Spanish, and French. The study is justified due to the importance of this industry in Brazil, as the country is an important textile producer and it possess the biggest textile chain in the West, ranging from planting cotton and making clothes, to retail and fashion shows (ABIT – ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL E DE CONFECÇÃO, 2023). The terms were extracted using a comparable oral language corpus build during the course of this work from the automatic transcription of 320 YouTube videos, 80 for each language. Once extracted, the terms were analyzed regarding their frequency in the corpus, their relevance in the fashion field, the contexts in which they appear, their definitions and the collocates that accompany them. Having collected that information, it was possible to identify equivalences between the languages, based on their contexts and definitions (GODOY, 2019). All this information was organized into a terminology dossier (PAVEL; NOLET, 2002) and afterwards transferred to terminology records (KRIEGER, 2004). This paper presents a cross-section of the 30 most frequent terms in the corpus and their 74 collocates. When the identification of equivalences was completed, 10 terms were found to have equivalences in all 4 languages, 20 had equivalences in up to 3 languages, 26 had equivalences in up to 2 languages and 45 had no equivalences.*
Keywords: *Terminology, Corpus Linguistics, Fashion, Multilingualism.*

Introdução

A indústria da moda é uma das maiores indústrias globais e compreende desde a plantação e extração de matéria prima, passando pela confecção de tecidos, até a construção, promoção e venda de produtos de vestuário, calçados, acessórios e joias (MARTINS et al., 2023). Segundo a ABIT (ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL E DE CONFECÇÃO, 2023), essa indústria emprega, no Brasil, cerca de 1,34 milhão de brasileiros

formalmente, sendo 60% de mão de obra feminina. O país possui também a maior cadeia têxtil completa do ocidente, no qual a indústria vai desde a plantação de algodão, até a confecção das peças, os desfiles de moda e o varejo. É uma indústria de quase 200 anos no país, que está entre os maiores produtores de denim e malha (ABIT - ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL E DE CONFECÇÃO, 2023).

Por estes motivos, globalmente, o Brasil é um importante produtor têxtil e inevitavelmente essa indústria se beneficia, para realizar essas trocas comerciais, do uso de outros idiomas, majoritariamente o inglês, o espanhol e o francês. Segundo Oliva e Barbosa (2023), o inglês e o espanhol são os dois idiomas obrigatórios nos currículos escolares brasileiros devido à sua importância internacional; o espanhol por ser um idioma presente em 160 países e com uma previsão de que, em aproximadamente 5 décadas, cerca de 500 milhões de pessoas falarão espanhol. Já o inglês é o idioma oficial do comércio internacional atualmente, e é:

> A língua principal do controle aéreo, do comércio exterior, dos eventos internacionais, da medicina, da diplomacia, das competições esportivas internacionais, da cultura pop, da informática, da internet, da cultura de massas, da ciência e da tecnologia. Dois terços dos cientistas escrevem em inglês. (OLIVA; BARBOSA, 2023)

O francês era um idioma curricular de caráter obrigatório até os anos 50 no Brasil que ainda possui uma importância significativa (OLIVA; BARBOSA, 2023). Além disso, o francês é o idioma que mais tem influência na área da moda, devido à importância cultural da França nesta indústria durante séculos. Boa parte do vocabulário da moda, em diversos idiomas, têm origem no francês (RAIMUNDO, 2003). Sendo assim, a criação de produtos terminológicos como dicionários e glossários bilíngues e/ou multilíngues, que tenham o português como idioma de partida e esses três outros idiomas como idiomas de chegada é de extrema importância. No entanto, quando pesquisamos por estes produtos, como glossários e dicionários, os resultados não são satisfatórios. É possível encontrar trabalhos como o realizado por Farias e Bezerra (2009), que se propõe à criação de um glossário multilíngue na área do vestuário, tendo o português como língua de partida e o inglês e francês como línguas de chegada, ou como o proposto por Raimundo (2003) que se propõe a criação de um glossário bilíngue em português e francês, ou ainda como o proposto por Fiasco (2018) que se propõe a criar um glossário em inglês, francês e italiano, tendo o inglês como língua de partida. Como se observa, uma investigação compreendendo a extração e análise para construção de um glossário multilíngue de moda ainda é escassa. Assim, emerge a necessidade da extração e análise de termos da área da moda em português, inglês, espanhol e francês para uma criação futura de um produto como um glossário devido à importância cultural desses idiomas, como

já apresentado anteriormente, para auxiliar tradutores, estudantes, pesquisadores e demais profissionais que trabalham na ou com a indústria da moda.

Portanto, este artigo tem como objetivo geral extrair e analisar termos da área da moda nos idiomas português, inglês, espanhol e francês.

Tal objetivo geral é constituído dos seguintes objetivos específicos:

- Criar um corpus multilíngue através da transcrição automática de vídeos do YouTube;
- Extrair e analisar candidatos a termo de forma semiautomática utilizando o software AntConc;
- Identificar equivalências dos candidatos a termo entre os diferentes idiomas propostos a partir das características identificadas na análise.

O presente trabalho busca contribuir para o estudo de Terminologia ao ampliar os estudos realizados com corpora orais que ainda não são muito difundidos e ao fornecer produtos com dados relevantes para consulta e estudo, tais como planilhas e tabelas, e uma metodologia que pode ser replicada ou discutida na extração e análise terminológica multilíngue. Busca contribuir com a indústria da moda ao fornecer termos importantes da indústria em 4 idiomas diferentes para serem utilizados por tradutores, estudantes e pessoas que trabalham em empresas com um contexto globalizado. Este trabalho também busca ampliar os conhecimentos adquiridos no curso de Línguas Estrangeiras Aplicadas - MSI da Universidade de Brasília ao aplicar de forma prática conteúdos teóricos vistos em disciplinas como Linguística de Corpus e Língua, Léxico e Terminologia, apresentando as possibilidades e as dificuldades no processo de extração e análise terminológica multilíngue baseadas em corpus.

A primeira seção do artigo apresenta a fundamentação teórica e está dividida em "Terminologia: extração e análise de termos" e "Linguística de corpus: etapas e critérios para a compilação de corpora". A segunda seção apresenta a metodologia utilizada neste estudo, na qual as informações apresentadas na fundamentação teórica são testadas de forma prática, esta seção está dividida em "Compilação dos corpora de estudo e referência" e "Extração de palavras-chave e identificação de terminologia". A terceira seção apresenta uma análise dos dados obtidos a partir da metodologia apresentada na seção 2. A quarta e última seção apresenta as considerações finais e os resultados da investigação, além das possibilidades para pesquisa futura.

1. Fundamentação teórica

1.1 Terminologia: extração e análise de termos

Nesta seção serão apresentadas todas as etapas que constituem o trabalho do terminólogo, desde a extração dos termos, até a sua análise e o seu armazenamento em dossiês e fichas terminológicas. Tal trabalho consiste, segundo Almeida e Correia (2008), na organização, no estudo e na apresentação de um vocabulário específico de uma ciência, arte, técnica ou atividade profissional, vocabulário este conhecido também como língua de especialidade. Sendo assim, trata-se de identificar termos de uma dada área; atestar o emprego desses termos; descrevê-los com clareza e concisão, aconselhando ou desaconselhando certos usos. É ainda realizado por meio da análise terminológica, que consiste em analisar textos de uma dada área e selecionar seus conceitos específicos que são designados por unidades terminológicas (PAVEL; NOLET, 2002).

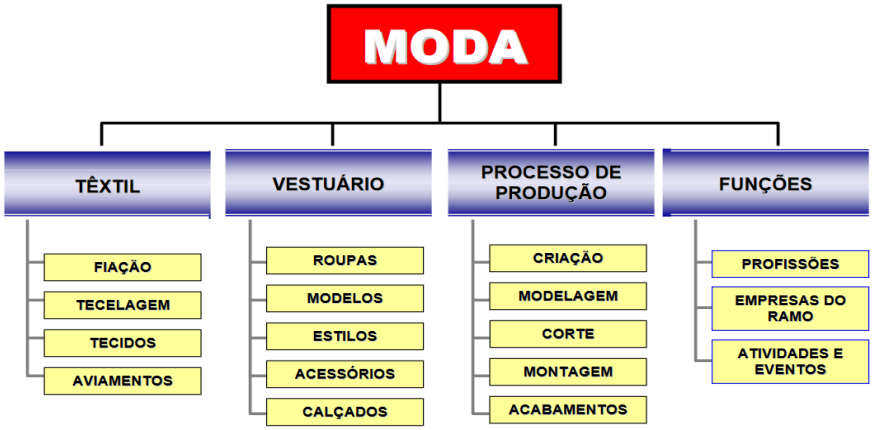
A Terminologia trata também da criação de glossários bilíngues ou multilíngues, que busca por equivalências entre idiomas diferentes. Segundo Welcker (2004), "O dicionário monolíngüe geralmente oferece definições, ao passo que o bilíngüe fornece sinônimos, mas na outra língua" (WELCKER, 2004, p. 194). Esses sinônimos, também chamados equivalentes, podem ser encontrados através do uso de corpora paralelos, cada texto com sua respectiva tradução, ou de corpora comparáveis, textos em idiomas diferentes que tratam de assuntos relacionados (PRESTES, 2015). Quando o pesquisador utiliza corpora comparáveis, é possível encontrar equivalências entre os idiomas através de ganchos terminológicos, como apresentado por Godoy (2019), nessa metodologia, os equivalentes são identificados através das características dos candidatos a termo nos contextos em que aparecem, principalmente as características que revelam natureza, finalidade e modo. O contexto identificado pode ser definitório, quando oferece características precisas do termo; pode ser explicativo, quando fornece apenas uma ideia da natureza do termo; ou pode ser associativo, quando relaciona o termo com outros termos que pertencem ao mesmo contexto, mas não o define claramente (GODOY, 2019). No caso da falta de equivalências, cabe ao terminólogo sugerir um novo termo, baseando-se nas regras de formação lexical do idioma. Cabe também ao terminólogo a constante atualização dos dados para que o trabalho se mantenha atual e coerente, isto pode ser feito através de modificações, supressões e acréscimos (PAVEL; NOLET, 2002).

O trabalho terminológico inicia-se pela criação de um banco de dados ou corpus, que será detalhado na seção "Linguística de Corpus: etapas e critérios para a compilação de corpora", e que é definido por Berber Sardinha (2004) como:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguística ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SARDINHA, 2004, p. 18)

O pesquisador deve então definir o campo temático da pesquisa, ou seja, dividir áreas e subáreas temáticas que englobam os termos (PAVEL; NOLET, 2002). A área da moda, assim como outras áreas do conhecimento, pode ser dividida em subáreas diferentes a depender do foco da pesquisa. Raimundo (2003) utiliza o organograma apresentado na figura 1 a seguir para dividir as subáreas dentro da moda. A subárea "têxtil" engloba a indústria têxtil, indo desde a criação do tecido (fiação, tecelagem), passando pelos tipos diferentes de tecidos, e chegando em aviamento, que é "tudo aquilo que vai na roupa, ficando nela permanentemente" (RAIMUNDO, 2003, p.73). A subárea "vestuário" diz respeito às peças do vestuário (roupas, calçados, acessórios), além dos modelos diferentes de cada peça e dos estilos, que são as tendências e as características do modo de vestir individual ou de grupos. A subárea "processo de produção" abrange a produção de cada peça do vestuário, desde o pensar a peça pelo estilista (criação), até as etapas mais manuais, feitas pelas modelistas (modelagem) e costureiras (corte, montagem e acabamento). Por fim, a subárea "funções" abarca os termos "ligado[s] ao trabalho, à execução de tarefas e ofícios" (RAIMUNDO, 2003, p. 76), por isso estão aí incluídos profissões, empresas, atividades e eventos.

Figura 1 - Organograma de áreas e subáreas de moda



Fonte: Raimundo, 2003.

Raimundo (2003) elaborou um glossário bilíngue e usou livros, periódicos e materiais da Biblioteca da Universidade Estadual de Londrina para fazer um levantamento bibliográfico

de onde os termos foram extraídos e analisados. O glossário foi finalizado com 270 termos em português e suas respectivas traduções em francês.

Farias (2003) divide os subdomínios da área da moda em: "1. tecido 2. padrão 3. cor 4. vestuário 5. estilo" (FARIAS, 2003, p. 73). O glossário construído por Farias e Bezerra (2009) utiliza 3 idiomas (português, francês e inglês), a pesquisa foi iniciada usando 3 dos subdomínios apresentados por Farias (2003), padrão, vestuário e estilo, mas foi posteriormente reduzido apenas ao subdomínio vestuário devido à falta de tempo, foram identificados 254 termos em português, com suas respectivas traduções em inglês e francês. Já Fiasco (2018) separa os termos de sua pesquisa em "clothes, accessories, textiles and materials, jewels, the fashion industry" (FIASCO, 2018, p. 3). Fiasco (2018) utilizou revistas de moda, como a Vogue e a Elle, em 3 idiomas diferentes (inglês, italiano e francês) para construir seu corpus e identificou 375 termos, totalizando 125 entradas em inglês, idioma de partida, com suas respectivas traduções.

Com o corpus, o campo temático e as subáreas definidos, o pesquisador pode iniciar a extração dos termos. O processo de extração de termos consiste na identificação da terminologia especializada dentro do corpus e na separação entre termos de domínio específico e palavras de língua geral. Este processo pode ser automatizado ou semiautomatizado e é estudado pela Terminologia Computacional, área que combina os avanços da Terminologia com o processamento de linguagem natural e a inteligência artificial para tratar grandes bancos de dados dentro da Terminologia (KRIEGER, 2004).

Uma das técnicas de extração de termos é através do uso de softwares de extração terminológica automática, como o ExATOlp (LOPES; FINATTO; CIULLA, 2015), o WordSmith e o AntConc (KADER; RICHTER, 2013), os quais usam uma metodologia que compara um corpus de estudo, criado com documentos específicos de uma dada área, e um corpus de referência, geralmente encontrado na internet e que contém palavras comuns da língua geral (ASSIS, 2020). Esta comparação exclui palavras com ocorrências similares nos dois corpora e destaca discrepâncias significativas, identificando as possíveis palavras-chave do corpus de estudo que serão analisadas a saber se são termos ou apenas palavras da língua geral.

A etapa seguinte à extração de termos, é a categorização. Pavel e Nolet (2002) sugerem a criação de um dossiê terminológico antes das fichas terminológicas. Esse dossiê é um documento que contém todos os candidatos a termo, que fornece uma visão macro da nomenclatura do trabalho, e no qual serão guardadas todas as provas textuais de cada conceito,

são elas definição, contexto, exemplos de uso, observações e referência. Essas provas textuais são definidas por Pavel e Nolet (2002) da seguinte forma:

> A definição enumera os traços semânticos que distinguem um conceito de todo os outros; o contexto é uma citação que ilustra a definição; os exemplos de uso mostram o funcionamento dos termos no discurso de especialidade; as notas ou observações precisam o uso dos termos no discurso; as referências indicam as fontes das provas textuais. (PAVEL; NOLET, 2002, p. XIX).

Como o dossiê é apenas um documento que organiza todos os termos, após a sua categorização, os termos são divididos em fichas terminológicas (PAVEL; NOLET, 2002). Cada termo possui uma ficha e ela constitui um "registro completo e organizado de informações referentes a um dado termo" (KRIEGER, 2004, p. 136). Segundo Krieger (2004) essas fichas são compostas de informações indispensáveis, definidas como "a fonte textual de coleta de um termo, segmentos de texto onde esse termo ocorre, seus contextos de uso, informações sobre variantes denominativas, sinônimos, construções recorrentes que o acompanham" (KRIEGER, 2004, p. 136) e informações operacionais, como "nome do responsável pela coleta, datas de registro e revisão" (KRIEGER, 2004, p. 136). Pavel e Nolet (2002) defendem que as seguintes informações devem estar presentes nas fichas terminológicas:

> "a que áreas temáticas pertence o conceito, as línguas às quais se circunscreve, os termos que designam o conceito em cada uma dessas línguas, a definição do conceito ou qualquer outro tipo de prova textual e as fontes que documentam essa informação." (PAVEL; NOLET, 2002, p. 9).

A Terminologia trabalha com o conceito Uninocional, ou seja, que cada termo designa um conceito, sendo assim, uma ficha terminológica não pode conter mais de um termo ou mais de um conceito, se for o caso de sinônimos, é necessário criar uma ficha terminológica para cada acepção (PAVEL; NOLET, 2002).

Outra informação que pode ser analisada em relação aos candidatos a termo é a pertinência dentro daquela nomenclatura. Nem sempre na fase de extração os candidatos a termo encontrados pertencem diretamente à área analisada, alguns termos não são do domínio *stricto sensu*, ou seja, possuem apenas uma relação com a área, mas é importante mencioná-los (KRIEGER, 2004). Segundo Krieger (2004), os termos identificados em um trabalho terminológico são separados por pertinência temática ou pertinência pragmática. A pertinência temática diz respeito aos termos que fazem parte daquele domínio *stricto sensu*, ou seja, eles compõem um núcleo estável que "carrega os traços distintivos que representam a especificidade da área" (KRIEGER, 2004, p. 138). São traços exclusivos do domínio, nítidos e constantes. Já os termos de pertinência pragmática estão relacionados ao domínio *lato sensu*, ou seja, termos de áreas correlatas e que necessitam de explicação devido ao contexto no qual aparecem.

Todas essas informações coletadas servirão para a criação de um produto que cumpra o objetivo final da pesquisa, levando em consideração o público-alvo do produto e o conhecimento que se pressupõe que esse público tenha. Os dados armazenados nas fichas terminológicas servem, por exemplo, para a criação de verbetes que compõem os glossários e os dicionários (KRIEGER, 2004). A terminologia e a nomenclatura de cada área possuem particularidades, áreas e subáreas e diferenças lexicais que precisam ser levadas em conta pelo terminólogo ao iniciar sua pesquisa, por isso é importante que o terminólogo ou alguém de sua equipe possua um conhecimento na área que será estudada (KRIEGER, 2004). A relação entre Terminologia e Linguística de Corpus não é recente. No entanto, a maioria dos trabalhos realizados atrelados a essas duas disciplinas são feitas a partir de corpora compostos de textos escritos. Os trabalhos realizados a partir de corpora orais ainda são escassos e pouco estudados, apesar de ser um campo muito grande, que apresenta uma parte linguística importante (BRUM-DE-PAULA; ESPINAR, 2002). Com os avanços tecnológicos, pouco a pouco as redes sociais e, principalmente, os vídeos do YouTube e os podcasts, substituem as revistas e os manuais de instruções, que serviam para a construção de corpora. Por isso, trabalhos que usam corpora orais são muito importantes e um campo que ainda precisa de muita pesquisa (BRUM-DE-PAULA; ESPINAR, 2002), e por este motivo a presente investigação se propõe à compilação e análise de corpora orais. Na próxima seção serão apresentadas as etapas e critérios para a compilação de corpora usados durante o trabalho terminológico.

1.2 Linguística de Corpus: Etapas e critérios para a compilação de corpora

De acordo com Almeida e Correia (2008) para a realização de um trabalho dentro da Terminologia, principalmente aquele que tem como objetivo a elaboração de um produto (dicionários, glossários, vocabulários), é fundamental a utilização de corpora, que é usado para "extração de candidatos a termo e suas formas variantes, elaboração da ontologia e a redação da definição terminológica." (ALMEIDA; CORREIA, 2008, p. 74). Segundo Berber Sardinha (2005), quanto maior o corpus de referência, maior será a quantidade de palavras-chave identificadas, o que influenciará diretamente na pesquisa realizada, por isso o corpus de referência deve ser pelo menos 2 vezes maior do que o corpus de estudo (SARDINHA, 2005).

Para essa coleta de dados existem alguns critérios e pré-requisitos os quais o pesquisador deve considerar. Como pré-requisito, o pesquisador deve considerar se os textos são autênticos, não tendo sido desenvolvidos para pesquisa, se os falantes ou escritores são nativos, se o

conteúdo foi escolhido criteriosamente e se é representativo dentro dos objetivos da pesquisa (SARDINHA, 2004).

Já os critérios que devem ser considerados pelo pesquisador são: modo; tempo; seleção; conteúdo; autoria; disposição interna; e finalidade. O modo diz respeito ao corpus ser falado ou escrito; o tempo indica se o corpus é sincrônico, diacrônico, contemporâneo ou histórico e estes são definidos por Sardinha (2004) como: "Sincrônico: compreende um período de tempo. Diacrônico: compreende vários períodos de tempo. Contemporâneo: representa o período de tempo corrente. Histórico: representa um período de tempo passado." (SARDINHA, 2004, p. 20). A seleção do corpus pode ser de amostragem, porção finita de textos ou variedades textuais que representam uma parte da língua; monitor, que pode ser aumentado para refletir o estado atual da língua; dinâmico, que permite o crescimento ou a diminuição; estático, que se opõe ao dinâmico; ou equilibrado, que possui quantidades semelhantes de gêneros textuais por exemplo. O conteúdo pode ser especializado, que utiliza textos específicos de um gênero ou registro; regional ou dialetal, textos que apresentam uma ou mais variedades sociolinguísticas; e/ou multilíngue, que inclui idiomas diferentes. A autoria separa os corpora em: de aprendiz, falantes que possuem aquele idioma como segunda língua, ou de língua nativa. A disposição interna diz respeito ao corpus ser paralelo, cada texto com sua respectiva tradução, ou alinhado, textos da mesma área, mas que não são traduções diretas. E, por fim, a finalidade do corpus indica se é um corpus de estudo, de referência, de treinamento ou de teste, definidos por Sardinha (2004) como: "De estudo: o corpus que se pretende descrever. De referência: usado para fins de contraste com o corpus de estudo. De treinamento ou de teste: construído para permitir o desenvolvimento de aplicações e ferramentas de análise (SARDINHA, 2004, p. 21).

A compilação do Corpus, segundo Almeida e Correia (2008), é composta de algumas etapas, a primeira delas é a delimitação do domínio, partindo para a seleção das fontes, que deve levar em consideração as publicações oficiais na área, o tipo de produto que será gerado a partir desse corpus e o público-alvo do produto. Após todas as fontes selecionadas, é necessário converter os documentos para a extensão *txt* e fazer uma limpeza de imagens, links ou outros elementos que podem atrapalhar o tratamento dos dados. Os documentos devem ser nomeados de maneira transparente, seguindo a seguinte ordem: meio de divulgação, gênero discursivo, fonte, data de publicação. Após nomeados, a última etapa é a anotação dos textos, que pode ser estrutural ou linguística, no entanto essa fase não é obrigatória, pois é possível realizar um trabalho terminológico com documentos não anotados (ALMEIDA; CORREIA, 2008).

Com o corpus compilado pode-se iniciar a análise dos dados, para isso existem diversos softwares, um destes softwares é o AntConc (KADER; RICHTER, 2013). Kader e Richter

(2013) apresentam as diversas ferramentas deste software, como KWIC ou *Key Words in Context*, *cluster*, *collocates*, *N-gram, Wordcloud* e *KeyWords*. A ferramenta KWIC ou *Key Words in Context* fornece o contexto no qual cada termo aparece. As ferramentas *cluster* e *collocates* fornecem listas de palavras que aparecem frequentemente com determinados termos, ou seja, palavras que formam um vocabulário específico de uma dada ciência ou arte (ALMEIDA; CORREIA, 2008). Na identificação de *collocates*, é possível que o pesquisador delimite o "*Window Span*", ou seja, a distância que uma palavra pode aparecer do termo, pois mesmo que uma palavra não esteja diretamente unida à outra, elas podem ser consideradas como *collocate* se aparecerem próximas de maneira significativa. É possível definir a distância entre essas palavras tanto para a direita, quanto para a esquerda. Já a ferramenta *KeyWords* gera uma lista de palavras-chave através da comparação de um corpus de estudo e de um corpus de referência. Os corpora de referência, também chamados de corpora geral, geralmente são monolíngues e apresentam uma variedade linguística de um dado idioma (VAUGHAN; O'KEEFFE, 2015). Para que essa comparação entre um corpus de estudo e um corpus de referência seja realizada, geralmente os pesquisadores utilizam listas de *stopwords*, essas listas contêm palavras que são insignificantes para a pesquisa e podem ser compostas de preposições, conjunções e artigos (KAUR; BUTTAR, 2018).

Outra informação que é levada em consideração na identificação de candidatos a termo é a quantidade de *tokens* e *types* dentro dos corpora. *Tokens* são "sequências individuais de caracteres que o reconhece como palavras individuais" (VAUGHAN; O'KEEFFE, 2015, p. 5, tradução própria)[1], dessas palavras individuais muitas são repetidas, portanto a informação *types* identifica as palavras distintas que aparecem nos textos. A diferença entre *tokens* e *types* pode ser mais bem entendida com o exemplo fornecido por Vaughan e O'Keeffe (2015) com a frase "you put your right leg in, your right leg out" que contém 10 *tokens* e 7 tipos diferentes de *types*, como demonstrado na figura 2 a seguir:

Figura 2 - Demonstração de tokens e types

Tokens									
You	put	your	right	leg	in,	your	right	leg	out
1	2	3	4	5	6	7	8	9	10
Types									
You	put	your	right	leg	in,	your	right	leg	out
1	2	3	4	5	6	3	4	5	7

Fonte: Vaughan e O'Keeffe, 2015

OBJ

A próxima seção apresentará a metodologia utilizada nesta pesquisa para a compilação de corpora, a extração e a análise dos dados coletados, seguindo as propostas apresentadas aqui na Fundamentação teórica.

2. Metodologia

O processo metodológico seguiu o apresentado pelo Manual de Terminologia (2002), que compreende, resumidamente: a criação de um banco de dados/corpus; a delimitação de um campo temático; o estabelecimento de uma árvore de conceitos; a extração de termos; a compilação dos dados extraídos em dossiês terminológicos, organizados posteriormente em fichas terminológicas; e o gerenciamento do conteúdo terminológico de acordo com a necessidade do usuário. O corpus foi criado a partir da transcrição automática de vídeos do YouTube, visando responder a seguinte pergunta: Seria possível extrair termos relevantes de uma área usando apenas transcrições automáticas e uma metodologia de extração de termos também semiautomatizada?

2.1 Compilação dos corpora de estudo e de referência

Nesta seção será apresentada a metodologia utilizada para a criação de corpora para serem usados na extração de termos e das provas textuais, essa metodologia segue as propostas apresentadas na Fundamentação Teórica discutidas nas seções anteriores. O corpus utilizado no presente trabalho foi criado a partir da transcrição de 320 vídeos do YouTube, já que a utilização de corpora de língua falada ainda é pouco realizada nos estudos da Linguística de Corpus em relação aos corpora cuja fonte advém de textos escritos (BRUM-DE-PAULA; ESPINAR, 2002). Como os vídeos são de uma mesma área, mas não são traduções diretas uns dos outros, o corpus é comparável e não paralelo. A seleção dos vídeos e dos falantes seguiram os critérios apresentados por Berber Sardinha (2004), sendo eles: modo: falado; tempo: sincrônico e contemporâneo, vídeos feitos entre 2018-2022; conteúdo: especializado, na área de moda, e multilíngue, em português, inglês, espanhol e francês, idiomas ofertados pelo curso de Línguas Estrangeiras Aplicadas - MSI da Universidade de Brasília; autoria: de língua nativa; disposição interna: alinhado; finalidade: de estudo; e seleção: dinâmico, é possível aumentar a quantidade de vídeos transcritos, e equilibrado, no final ficaram 80 vídeos para cada idioma, procurou-se equilibrar também a quantidade de canais diferentes para cada idioma e a quantidade de vídeos dentro de cada subárea temática diferente (Tendências, Tutoriais e História da Moda), como apresentado nas Tabelas 1 e 2 abaixo.

Tabela 1 - Listagem de canais do YouTube separados por idioma e áreas temáticas

Idioma	Tendências	Tutoriais	História da Moda	**TOTAL (idiomas)**
TOTAL (subáreas)	**13**	**13**	**14**	**40**
Português	3	3	4	**10**
Inglês	3	4	3	**10**
Espanhol	4	3	3	**10**
Francês	3	3	4	**10**

Fonte: Da autora, 2023.

Tabela 2 - Listagem da quantidade de vídeos separados por idioma e áreas temáticas

Idioma	Tendências	Tutoriais	História da Moda	**TOTAL (idiomas)**
TOTAL (subáreas)	**113**	**117**	**90**	**320**
Português	21	27	32	80
Inglês	30	37	13	80
Espanhol	27	27	26	80
Francês	35	26	19	80

Fonte: Da autora, 2023.

A escolha dos canais foi realizada através de pesquisa no YouTube, pois esta é a maior plataforma de conteúdo audiovisual mundialmente e, segundo Nagumo, Teles e Silva (2020), no Brasil é a "plataforma mais utilizada para fins de conhecimento" (NAGUMO; TELES; SILVA, 2020, p.3). Dentro da plataforma buscou-se as palavras-chave "moda, costura, história da moda, tendências de moda", termos mais gerais da área da moda, e suas respectivas traduções para os outros 3 idiomas, e usando a ferramenta de filtro, onde foi selecionada a opção "canal". A partir daí, os canais que apareceram foram analisados para saber se os vídeos publicados estavam dentro dos critérios de tamanho (15-30 minutos) e data (2018-2022). Esses critérios foram utilizados pois considerou-se que vídeos com menos de 15 minutos teriam pouco vocabulário que poderia ser utilizado para a investigação e vídeos com mais de 30 minutos apresentavam mais problemas no momento da transcrição automática. A data foi definida entre 2018-2022 pois representa um recorte atual dos idiomas estudados. Os canais foram verificados para saber se tratava apenas de moda e se em suas descrições explicitava-se isso e se os youtubers tinham alguma formação na área ou exerciam alguma função dentro da indústria da

moda (consultores de imagem, designers, costureiras, jornalistas). Também buscou-se selecionar criadores de conteúdo de regiões e países diferentes para ter uma variedade linguística. Uma amostra da planilha (Apêndice A) criada a fim de organizar os dados sobre cada youtuber está disposta em Apêndice.

Com todos os canais e vídeos selecionados, uma planilha no Google Planilhas (Apêndice B) foi criada dividida entre os quatro idiomas que serão utilizados no glossário: português, espanhol, francês e inglês, onde foram armazenadas as seguintes informações: a área temática de cada canal, o nome do canal, o nome do vídeo, a duração do vídeo, a data de publicação, o status da transcrição (se concluída ou não) e o link para o vídeo.

A coleta das transcrições dos vídeos foi feita de maneira automática usando a ferramenta de "Digitação por Voz" do Google Docs alinhada ao Cabo Virtual VB, disponível para download no link (https://vb-audio.com/Cable/). Este software permite que todo o som identificado pelo input, som transmitido pelo computador, seja diretamente encaminhado ao output do software, e não ao alto-falante do computador, por isso quando a ferramenta de Digitação por Voz é ativada, ela consegue identificar as palavras sem ruídos externos ao computador. O que nos fornece uma transcrição de maior qualidade.

Apesar de ter facilitado muito o trabalho, o processo de transcrição automática apresentou alguns problemas. Primeiro, dependendo de como a pessoa fala no vídeo, o tom da voz, ou se tem algum som de fundo, essa ferramenta muitas vezes não consegue identificar o que está sendo falado. Ela também apresenta muitos problemas para identificar o que está sendo dito quando tem mais de uma pessoa no vídeo, então a maioria dos vídeos de colaboradores ou os canais que possuíam mais de um criador de conteúdo foram retirados, pois a transcrição sempre pausava assim que uma pessoa diferente começava a falar. Outro problema apresentado foi em relação aos anúncios que aparecem ao decorrer do vídeo e que fazem com que ocorra uma falha que pausa a transcrição. Por causa desses problemas, o processo acaba sendo semiautomático, já que é necessário estar sempre atento ao computador e reiniciar o programa caso haja alguma falha.

Durante o processo de extração e análise dos termos, outros problemas foram identificados. A falta de pontuação, por exemplo, pode identificar *collocates* que na verdade são palavras em frases diferentes, que deveriam estar separadas por vírgula ou ponto final. E por fim, foram identificados problemas na ortografia das palavras que podem causar dificuldades no momento de extração de termos e *collocates*. Como é o caso do exemplo disponível no Apêndice C, aqui os *collocates* da palavra-chave em francês "*fil*" foram identificados, "*fil de fronce*" é um termo da costura que caracteriza o fio usado para franzir os

tecidos, no entanto, metade das transcrições identificaram "*fronce*" como "*france*", criando assim o *collocate* "*fil de France*" que não tem nenhum significado em francês.

Este problema também foi identificado nos outros idiomas, como é o caso do sinônimo de manga bufante em espanhol que é "*manga abullonada*", mas que foi também transcrito como "*manga bullonada*". Ou o caso de "pence" em português que foi transcrito muitas vezes como "pense", o que causa ainda mais confusão no momento de identificar corretamente a frequência dos termos, já que "pense" também pode ser o imperativo do verbo pensar. Outro caso foi com o termo "*pantalón simil cuero*" (Apêndice D) em espanhol, termo que identifica um tipo de couro. No momento de identificar os *collocate* de "*pantalón*", "*cuero*" aparecia com uma frequência alta, no entanto, ao usar a ferramenta "*Key Words in Context* - KWIC" que mostra o contexto daquela ocorrência, é possível notar que a ferramenta de transcrição automática transcreveu "*simil cuero*" como "*sin el cuero*", "*sin mil cuero*" e até mesmo "*civil cuero*".

Após realizada todas as transcrições, cada documento foi baixado no formato txt e nomeado. O processo de nomeação dos arquivos é uma parte muito importante pois "confere organização ao conjunto de textos compilados" (ALMEIDA; CORREIA, 2008, p. 85). Almeida e Correia (2008) sugere que a nomeação siga a seguinte ordem: meio de divulgação, gênero discursivo, fonte, data de publicação. Para se adequar às necessidades do presente trabalho, que usa vídeos como fonte para o corpus e não textos, a ordem foi adaptada da seguinte maneira: idioma, meio de divulgação, área temática, nome do canal, título do vídeo e data de publicação. As três primeiras informações são apresentadas com as seguintes siglas: idioma: PT (português); EN (inglês); ES (espanhol); FR (francês); meio de divulgação: YT (YouTube); área temática: TE (tendências); TU (tutoriais); HM (história da moda), como apresentado no Apêndice E.

Como citado anteriormente, o corpus criado para este trabalho será comparado com corpora de referência usando o programa AntConc para encontrar palavras-chave. Para cada idioma proposto na pesquisa, um corpus de referência foi encontrado. Os corpora de referência usados foram o BNC Baby para o inglês (OXFORD TEXT ARCHIVE, 2007), o CREA para o espanhol (REAL ACADEMIA ESPAÑOLA, 2021), o Corpus Brasileiro para o português (PROJETO AC/DC: CORPO CORPUS BRASILEIRO, 2021), e o Corpus "Mixed" de 2009 da *Leipzig Corpora Collection* para o francês (WORTSCHATZ LEIPZIG, 2009). O corpus BNC Baby é um recorte do corpus BNC que foi desenvolvido pela Universidade de Oxford e contém textos escritos e falados dos gêneros acadêmico, ficção, periódicos e conversação na variedade britânica do inglês. O corpus CREA foi desenvolvido pela *Real Academia Española* e contém textos escritos e falados de diferentes países falantes de espanhol como Argentina, Chile, Costa

Rica, Bolívia e Colômbia, os textos foram retirados de livros, periódicos, revistas, miscelâneas e produções orais. O Corpus Brasileiro foi desenvolvido por Berber Sardinha e equipe, e contém textos falados e escritos na variedade brasileira do português nos seguintes gêneros textuais: acadêmico, cinema e tv, educação, enciclopédia, esporte, informática, jornalismo, legislação, literatura, medicina, política, religião e técnico. O corpus "Mixed" em francês foi desenvolvido pela Leipzig Corpora Collection e é parte de um projeto que desenvolveu um conjunto de corpora comparáveis em diversos idiomas usando a internet para coletar os textos em sites como Wikipedia, periódicos online, páginas da web e FindLinks. Os corpora de referência foram selecionados levando em consideração que eles deveriam ser pelo menos 2 vezes maiores do que os corpora de estudo na quantidade de *tokens*, como proposto por Sardinha (2005). Neste trabalho usou-se também listas de *stopwords* que foram criadas a partir da junção das listas disponíveis nos links no Quadro 1, que contém preposições, conjunções, artigos e os verbos mais comuns de cada idioma. O Quadro 1 abaixo também apresenta os links para cada corpus de referência e a quantidade de *tokens* e *types* de cada corpus.

Quadro 1 - Listagem dos Corpora e das listas de *StopWords* usadas como Referência

Idioma	Corpus de referência	*Token*	*Type*	Listas de StopWords
Português	https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS	5.503.508	2.944.572	https://www.linguateca.pt/chave/stopwords/folha.MF300.txt https://www.ranks.nl/stopwords/portuguese https://gist.githubusercontent.com/alopes/5358189/raw/2107d809cca6b83ce3d8e04dbd9463283025284f/stopwords.txt
Inglês	https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2553	4.022.352	83.548	https://www.ranks.nl/stopwords
Espanhol	http://corpus.rae.es/lfrecuencias.html	743.435	733.743	https://www.ranks.nl/stopwords/spanish https://gist.github.com/cr0wg4n/78554c5d0afa9944d2fa3a4435d83a57#file-spanish-stop-words-txt
Francês	https://wortschatz.uni-leipzig.de/en/download/French	678.144	374.036	https://www.ranks.nl/stopwords/french https://www.destrucsaweb.com/ressources/liste-stop-words-francais.txt

				https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt

Fonte: Da autora, 2023.

Com os corpora de estudo prontos e nomeados e os corpora de referência e as listas de *stopwords* compilados, é possível passar então para a extração dos termos que será apresentada na seção abaixo.

2.2 Extração de palavras-chave e identificação de terminologia

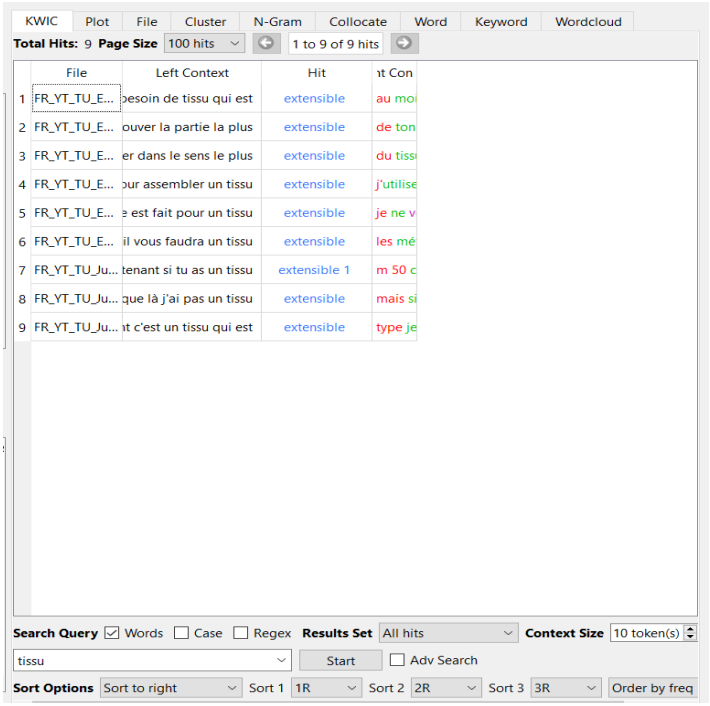
Com os corpora compilados e nomeados, deu-se início à extração de termos, que foi realizada separadamente para cada idioma e foi iniciada pelo idioma de partida, o português. No software AntConc, o corpus de estudo foi selecionado em "*Target Corpus*" e o corpus de referência, já mencionado, em "*Reference Corpus*", a lista de stopwords também foi adicionada em "*Tool Filters*". A partir disto a lista de palavras-chave foi gerada através da ferramenta *KeyWords* e apresentou 1606 palavras-chave no total, muitas palavras-chave não pertencem à área da moda, sendo assim a extração de termos se torna semiautomática pois o terminólogo precisa realizar uma seleção manual (ALMEIDA; CORREIA, 2008). Devido à grande quantidade de palavras-chave gerada e o limite de tempo para a finalização da pesquisa, para esta seleção manual, apenas as palavras-chave que tinham uma frequência de 20 ou mais ocorrências no corpus foram selecionadas, diminuindo assim a quantidade de palavras-chave a serem analisadas para 726. Destas 726, foram extraídas as que pertenciam à área de moda, finalizando a extração com 178 candidatos a termo para o idioma português. O inglês apresentou 918 palavras-chave no total, 451 palavras-chave que tinham uma frequência de 20 ou mais ocorrências e 208 candidatos a termo da área. O espanhol apresentou 1494 palavras-chave total, 759 palavras-chave que tinham uma frequência de 20 ou mais ocorrências e 240 candidatos a termo da área. Por último, o francês apresentou 849 palavras-chave no total, 693 palavras-chave que tinham uma frequência de 20 ou mais ocorrências e 199 candidatos a termo da área de moda. A relação entre as palavras-chave dos idiomas está disposta no gráfico 2 na seção Análise. O Apêndice F mostra a lista de palavras-chave gerada pelo programa em português.

Todos os candidatos a termo foram adicionados na planilha "Dossiê terminológico" (Apêndice G), que consiste em uma adaptação das informações necessárias em um dossiê, apresentada por Pavel e Nolet (2002). A planilha apresenta as seguintes colunas: pertinência,

subárea temática, palavras-chave, *collocates*, frequência no corpus, definição, fonte da definição, contexto e fonte do contexto.

Após isto, os candidatos a termo foram analisados quanto aos seus *collocates*, estruturas que acompanham o termo significativamente. O *Window Span* selecionado para encontrar os *collocate* de cada candidato a termo foi de 3 para a esquerda e 3 para a direita, pois 4 abrangia palavras que estavam muito distantes dos termos e 2 não identificava alguns termos com preposições e conjunções que os mantinham em uma distância maior, como é o caso da Figura 3 abaixo, na qual o termo "*tissu extensible*" em francês também aparece como "*tissu qui est extensible*". De cada *collocate* identificado, realizou-se uma seleção manual para identificar os que pertenciam à área da moda.

Figura 3 - Exemplo de collocates que não aparecem exatamente um ao lado do outro



Fonte: Da autora, 2023.

Usando o software AntConc, a frequência de cada termo e cada *collocate* foi identificada e adicionada na coluna "Frequência no Corpus" na planilha "Dossiê e equivalentes". Com todos os candidatos a termo e *collocates* dispostos na planilha, iniciou-se a categorização de cada termo. Devido à grande quantidade de candidatos a termo identificados na extração, viu-se a necessidade de fazer um recorte dos 30 termos com maior frequência em

cada idioma, apresentados no quadro 2 abaixo, para continuar a coleta de informações, o restante dos termos continuará dispostos na planilha para pesquisa futura.

Quadro 2 - 30 termos mais frequentes no corpus em cada idioma

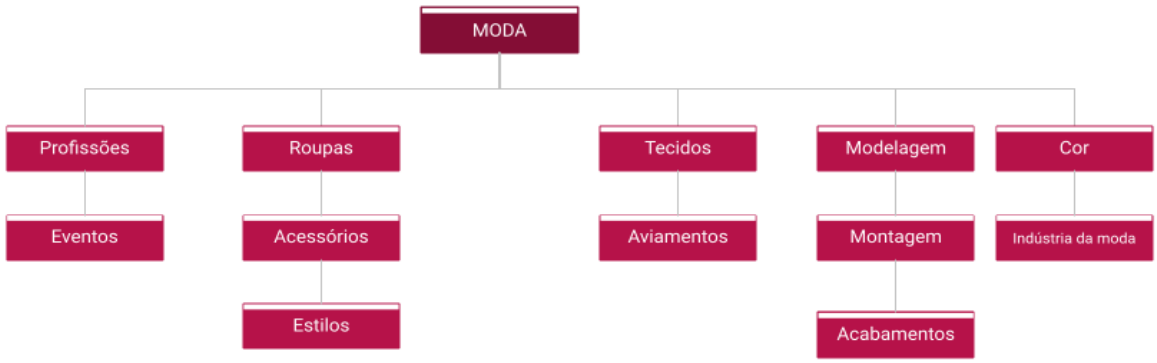
Português		Inglês		Espanhol		Francês	
Termos	Frequência	Termos	Frequência	Termos	Frequência	Termos	Frequência
costura	565	dress	568	moda	627	coup	655
vestido	502	fabric	451	costura	536	mode	527
moda	501	top	437	prendas	488	couture	438
frente	463	wear	417	color	399	robe	390
tecido	398	pieces	404	estilo	350	côté	383
peça	336	skirt	390	colores	324	tendance	375
costas	331	pattern	354	línea	315	tissu	351
cintura	322	fashion	346	manga	315	pièces	348
saia	304	style	315	vestido	305	porter	347
calça	300	front	303	tela	305	endroit	322
passar	275	stitch	293	piezas	300	tenue	273
estilo	269	wearing	292	pantalón	275	taille	258
blusa	253	side	270	prenda	215	collection	245
look	220	cut	264	negro	206	corps	231
cortar	217	sewing	244	escote	193	pantalon	214
manga	215	piece	230	pieza	190	gros	201
baixo	215	color	228	invierno	186	sac	189
lateral	201	black	225	blusa	184	jupe	186
decote	190	outfit	214	tendencia	184	top	167
branco	187	seam	205	tendencias	182	porte	167
bolsa	177	white	200	falda	177	défilé	165
alta	175	trend	179	cuerpo	174	col	163
linha	174	edge	158	pasar	170	coudre	162
corpo	171	jacket	157	ropa	169	passer	162
peças	167	fit	154	verano	164	saison	157
cor	163	lining	154	delantero	161	vêtements	155
cores	162	jeans	151	colección	160	couleur	148

costurar	161	denim	149	cuello	156	blanc	147
modelagem	156	pants	149	temporada	155	forme	147
ombro	156	bag	148	otoño	155	tendances	146

Fonte: Da autora, 2023.

Os termos acima foram primeiramente categorizados quanto à pertinência, se temática ou pragmática, de acordo com Krieger (2004) e depois quanto à subárea que pertenciam. As subáreas temáticas utilizadas foram adaptadas do organograma criado por Raimundo (2003) apresentado na Fundamentação Teórica, os termos aqui encontrados pertencem a 11 das subáreas presente no organograma, foi acrescido a elas a subárea "cor" apresentada por Farias (2003) e indústria da moda apresentada por Fiasco (2018) para abranger os candidatos a termo identificados. Todas as subáreas utilizadas neste trabalho estão dispostas no organograma abaixo:

Figura 4 - Organograma de subáreas temáticas da moda



Fonte: Da autora, 2023.

No final da pesquisa, quando todos os equivalentes entre os idiomas diferentes foram encontrados, a quantidade de termos por subárea ficou dividida da seguinte forma:

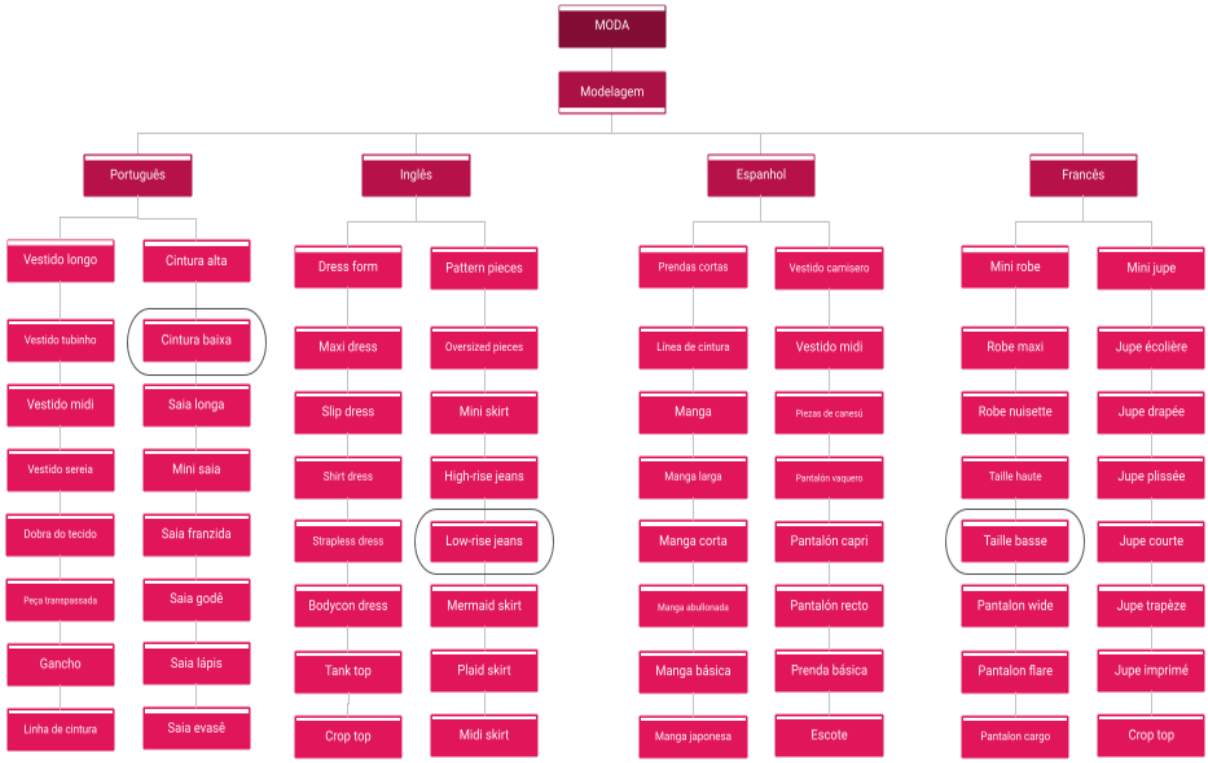
- Acabamentos: 0
- Acessórios: 3
- Aviamentos: 1
- Cor: 6
- Estilo: 3

- Indústria da moda: 2
- Modelagem: 21
- Montagem: 6
- Profissões: 0
- Roupas: 8
- Tecido: 5

Depois iniciou-se a identificação da definição de cada palavra-chave e dos *collocates*, a pesquisa foi feita usando glossários e dicionários já existentes e para cada definição foi adicionada também sua respectiva fonte. Após isto os contextos de cada candidato a termo foram identificados usando o corpus de estudo criado neste trabalho através da ferramenta KWIC - "*Key Words in Context*". As palavras-chave e os *collocates* para os quais não foi possível encontrar definições, foram desconsiderados como termo. Estas informações foram organizadas na planilha "Dossiê e equivalentes" disponível no Apêndice G.

Ao finalizar a coleta das definições e dos contextos, iniciou-se a identificação de equivalentes entre os idiomas. A metodologia utilizada foi a proposta por Godoy (2019), que usa a comparação de contextos e definições para identificar sinônimos em outros idiomas. Primeiro criou-se organogramas separados segundo cada subárea temática, em cada organograma são apresentados os candidatos a termo de cada subárea nos 4 idiomas diferentes usados neste estudo, sendo assim o equivalente para cada termo necessariamente estará dentro do mesmo organograma, pois eles pertencem a mesma subárea. Abaixo está o organograma da subárea "Modelagem", para exemplificar o processo de identificação de equivalentes, o termo "cintura baixa" será apresentado.

Figura 5 - Organograma da subárea modelagem contendo os candidatos a termo nos 4 idiomas diferentes



Fonte: Da autora, 2023.

Pode-se identificar no organograma acima o candidato a termo “cintura baixa” destacado. Para encontrar seus equivalentes é necessário comparar sua definição e contextos com os candidatos a termo que também aparecem nesse organograma, mas nos outros idiomas. A definição encontrada para esse candidato a termo, segundo o Dicionário do denim (LEVI’S® BRASIL, 2023), foi: “Jeans ou calças que ficam abaixo da cintura, próximos ao quadril. Um look casual e relaxado.”, e os contextos identificados no corpus foram:

1. Uma blusa de uma manga só, também característica dos anos 80, aparece mais uma vez com a calça de cintura baixa
2. o fato é que a cintura baixa aos pouquinhos está retornando, a gente tem visto ela com muito mais força do que anos atrás

Essa definição e esses contextos foram comparados aos outros candidatos a termo dentro do organograma nos outros idiomas, para o inglês foi possível identificar que o candidato a termo que mais se aproxima de “cintura baixa” é “*low-rise*”, definido pelo glossário de terminologia de denim (TRILOGY, 2023) como: “Low-rise jeans will rest on the hip bones.

Low-rise jean styles such as boyfriend jeans offer a more relaxed, casual look.", já os contextos identificados para o inglês foram:

1. I definitely see coming back in style, especially with low-rise jeans kind of being popular
2. which is being echoed in the mainstream with low-rise pants and skirts coming back in full force

Já na coluna em espanhol, não foi possível encontrar nenhum candidato a termo que poderia ser considerado um sinônimo para cintura baixa. Por fim, no francês, o candidato a termo que pode ser considerado um sinônimo para "cintura baixa" é "*taille basse*", que foi definido pelo dicionário da moda do Tabloïde Mode (2017) como: "Coupe particulière d'un pantalon, jean ou jupe, dont la taille se positionne plus sur les hanches, à environ 5 cm au-dessous du nombril". Os contextos identificados foram:

1. [...] prédit de plus en plus le retour du taille basse
2. Si tu souhaites l'utiliser pour une taille haute ou une taille basse, il faut prendre en compte la longueur en fait

É possível identificar similaridades tanto na definição fornecida por cada glossário/dicionário, quanto nos contextos apresentados, por exemplo, no contexto de todos os idiomas aparece a informação de que a cintura baixa está voltando à moda. Neste caso o contexto identificado é explicativo, pois fornece uma ideia da natureza do termo (GODOY, 2019).

Outro exemplo da identificação de equivalentes é o termo "alta-costura" presente no organograma da subárea "Estilo" (Apêndice H). A definição encontrada para "alta-costura" segundo o glossário de moda do blog Oscar (SIMÕES, 2023) é "são peças reconhecidas pela comissão sediada no Ministério da Indústria Francesa como roupas de luxo". Os contextos identificados foram:

1. As revistas femininas tinham um papel crucial ao ditar as tendências e mostrar com exclusividade os lançamentos da alta-costura
2. As casas de alta costura viveram talvez a sua grande Era de ouro nos anos 20
3. Ele recebeu a parte criativa uma grande *maison* francesa, de uma antiga casa de alta costura, então é um grande prestígio para ele e é também uma grande responsabilidade

Para o inglês, não foi possível encontrar dentro dos candidatos a termo nenhum sinônimo que se aproximasse da definição e dos contextos desse termo, sendo assim, para este termo a parte em inglês foi desconsiderada. Para o espanhol, o candidato a termo que se aproxima tanto da definição quanto dos contextos é "*alta costura*" definido pelo dicionário da moda da Audaces (2022) como "*T*érmino que se refiere a la ropa de alto lujo. Una pieza solo

puede recibir este reconocimiento si así lo decide el Ministerio de Industria francés.", e que apresentou os seguintes contextos:

1. Su primera colección de alta costura debutó 2 años después que la firma participó en París como visitante
2. cambiaría totalmente la visión de la moda, inicialmente su clientela de alta costura rechazó la idea, ya que era muy moderno
3. Hasta el verano de 1940, Elsa luchó por mantener su casa de alta costura para conservar el mayor número posible de puestos de trabajo e incluso lanzó un perfume

No idioma francês, o termo que foi considerado como sinônimo de "alta-costura" é "*haute couture*" definido pelo dicionário da moda do *Tabloïde Mode* (TABLOÏDE MODE, 2017) como "Désigne le domaine de la création de vêtements de luxe, conçus par les grands couturiers, et répondant à un certain nombre de critères notamment pour la qualité du travail produit et des professionnels employés à cet effet, et la présentation des modèles au cours de défilé à chaque saison.", e que apresentou os seguintes contextos:

1. Cette collection de haute couture était sa première collection de haute-couture, mais aussi c'était les 75 ans de la maison Balmain donc c'était un espèce d'anniversaire
2. on commence avec la mode qui se met au digital, notamment avec les collections haute-couture ces derniers temps
3. L'émergence du prêt-à-porter qui va faire que plein plein de maisons de haute couture vont mettre clé sous la porte, notamment Balenciaga en 1968

Aqui, da mesma forma, é possível comparar as definições oferecidas pelos glossários/dicionários, os três termos dizem respeito às roupas de luxo que seguem critérios estabelecidos pelo Ministério da Indústria Francesa. E neste caso, os contextos são associativos (GODOY, 2019), pois é possível relacionar o léxico apresentado em cada contexto, aparecem nos três idiomas palavras como coleções de alta-costura e casa de alta-costura.

Todos os candidatos a termo foram analisados usando a metodologia apresentada acima para a identificação de contextos, partindo sempre do idioma português, por isso, apenas os candidatos a termo que apareceram em português foram considerados. Dos 104 candidatos a termo identificados em português (palavras-chave e *collocates*), 10 apresentaram equivalências nos 4 idiomas, 20 apresentaram equivalências em até 3 idiomas, 26 apresentaram equivalências em até 2 idiomas e 45 não apresentaram equivalências. Todos os termos com seus equivalentes foram organizados na planilha "Dossiê e equivalentes", já apresentada aqui nesta seção, na pasta de trabalho "Entrada principal", como demonstrado no Apêndice I.

Ao finalizar a identificação dos equivalentes, iniciou-se a criação das fichas terminológicas. Como apresentado na seção de Fundamentação Teórica, as fichas terminológicas sistematizam e sintetizam todas essas informações coletadas sobre cada

candidato a termo e foram criadas baseando-se em Pavel e Nolet (2002) e Krieger (2004). As fichas foram criadas usando o Google Docs, cada documento contém as informações dos quatro idiomas utilizados nesta pesquisa e contempla as seguintes linhas: subárea temática; idioma; vídeos em que aparece; frequência no corpus; definição; fontes e contexto. Para cada termo criou-se uma ficha terminológica, resultando em 56 fichas terminológicas, ou seja, 56 termos que apresentam equivalências. Abaixo está a ficha criada para o termo "alta-costura".

Figura 6 - Ficha terminológica do termo "alta-costura"

Subárea temática: Estilo	
Idioma: Português	**Alta-costura**
Vídeos em que aparece:	13/80
Frequência no corpus:	55
Definição:	são peças reconhecidas pela comissão sediada no Ministério da Indústria Francesa como roupas de luxo.
Fontes:	https://blog.oscarcalcados.com.br/glossario-de-moda/
Contexto:	1. As revistas femininas tinham um papel crucial ao ditar as tendências e mostrar com exclusividade os lançamentos da alta-costura 2. As casas de alta costura viveram talvez a sua grande era de ouro nos anos 20 3. Ele recebeu a parte criativa uma grande maison francesa, de uma antiga casa de alta costura, então é um grande prestígio para ele e é também uma grande responsabilidade
Idioma: Inglês	
Vídeos em que aparece:	
Frequência no corpus:	
Definição:	
Fontes:	
Contexto:	
Idioma: Espanhol	**Alta costura**
Vídeos em que aparece:	13/80
Frequência no corpus:	111
Definição:	Término que se refiere a la ropa de alto lujo. Una pieza solo puede recibir este reconocimiento si así lo decide el Ministerio de Industria francés.
Fontes:	https://audaces.com/es/blog/diccionario-moda
Contexto:	1. Su primera colección de alta costura debutó 2 años después que la firma participó en París como visitante 2. cambiaría totalmente la visión de la moda, inicialmente su clientela de alta costura rechazó la idea, ya que era muy moderno 3. Hasta el verano de 1940, Elsa luchó por mantener su casa de alta costura para conservar el mayor número posible de puestos de trabajo e incluso lanzó un perfume

Idioma: Francês	Haute-couture
Vídeos em que aparece:	10/80
Frequência no corpus:	44
Definição:	Désigne le domaine de la création de vêtements de luxe, conçus par les grands couturiers, et répondant à un certain nombre de critères notamment pour la qualité du travail produit et des professionnels employés à cet effet, et la présentation des modèles au cours de défilé à chaque saison.
Fontes:	https://tabloidemode.media/lexique-de-la-mode/
Contexto:	1. Cette collection de haute couture était sa première collection de haute-couture, mais aussi c'était les 75 ans de la maison Balmain donc c'était un espèce d'anniversaire 2. on commence avec la mode qui se met au digital, notamment avec les collections haute-couture ces derniers temps 3. L'émergence du prêt-à-porter qui va faire que plein plein de maisons de haute couture vont mettre clé sous la porte, notamment Balenciaga en 1968

Legenda: a) Ficha terminológica em português, inglês e espanhol
b) Continuação da ficha terminológica em francês
Fonte: Da autora, 2023.

Como apresentado acima, nenhum sinônimo em inglês foi encontrado para esse termo, por este motivo a seção do idioma inglês ficou vazia, podendo ser preenchida com pesquisa futura que identifique algum equivalente. Partindo de todas as informações sobre os termos presentes nas fichas terminológicas e todos os dados coletados durante a pesquisa, a próxima seção traz uma análise desses dados.

3. Análise

Nesta seção serão apresentados a análise feita sobre os dados e os resultados obtidos. O método utilizado para a análise foi o método misto que combina análise quantitativa e qualitativa para obter resultados satisfatórios (TIMANS; WOUTERS; HEILBRON, 2019). Como citado na seção de Metodologia, os corpora de estudo foram criados a partir da transcrição automática de 320 vídeos, sendo 80 vídeos para cada idioma. Na tabela abaixo é possível identificar que os corpora de estudo, constituídos da mesma quantidade de vídeos e com tempo de duração parecido, resultaram em uma quantidade de *tokens* e *types* muito próxima. Já os corpora de referência encontrados na Internet, apesar de satisfazerem o critério de serem pelo menos 2 vezes maior do que os corpora de estudo, como sugerido por Berber Sardinha (2005), acabaram tendo uma quantidade bem diversa de *tokens* e *types*. A quantidade de *types* influenciou a quantidade de palavras-chave obtida para cada corpus, como apresentado no Gráfico 1 abaixo. O corpus de referência do inglês e do português possuem uma quantidade similar de *tokens*, no entanto, por apresentarem uma quantidade de *types* diferente, o inglês apresentou uma quantidade bem menor de palavras-chave. O francês apresentou a menor

quantidade de palavras-chave como já era esperado, pois o seu corpus de referência tem menos *tokens* e menos types do que os outros.

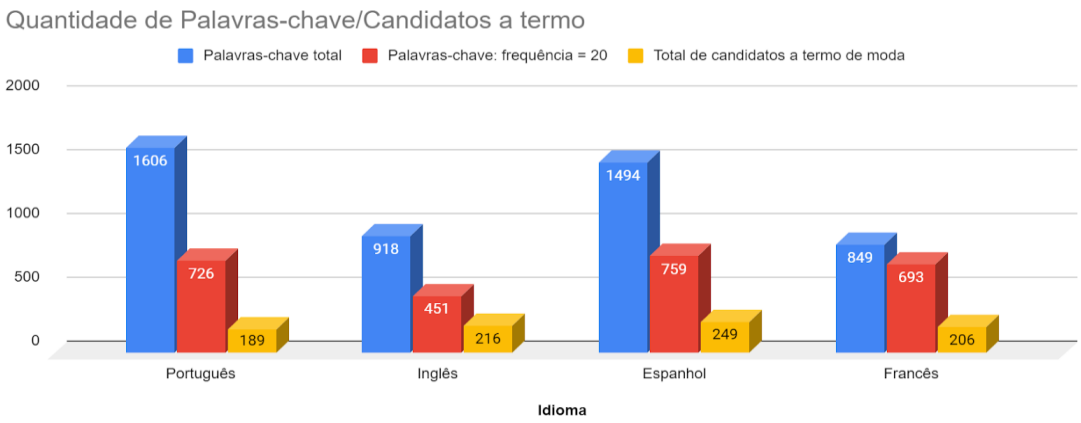
Tabela 3 - Relação de *tokens* e *types* dos Corpora de Estudo e dos Corpora de Referência

Idioma	Corpus de Estudo				Corpus de Referência			
	Tokens	%	*Types*	%	*Tokens*	%	*Types*	%
TOTAL	**1.013.821**	**100,0**	**53.893**	**100,0**	**10.947.439**	**100,0**	**4.135.899**	**100,0**
Português	237.972	23,5	14.027	26,0	5.503.508	50,3	2.944.572	71,2
Inglês	269.938	26,6	11.873	22,0	4.022.352	36,7	83.548	2,0
Espanhol	232.163	22,9	15.741	29,2	743.435	6,8	733.743	17,7
Francês	273.748	27,0	12.252	22,7	678.144	6,2	374.036	9,0

Fonte: Da autora, 2023.

Como já mencionado na seção de Metodologia, a extração dos termos foi realizada a partir da ferramenta *KeyWords* do software AntConc, após extraídas todas as palavras-chave, devido a grande quantidade de palavras-chave identificadas, a pesquisa foi limitada até os termos que tiveram uma frequência de 20 vezes ou mais no Corpus de Estudo (Target Corpus), as demais palavras-chave ficarão como extensão para pesquisa futura. A partir desses termos, foram identificados aqueles que eram da área da moda. O gráfico a seguir indica como ficaram estas relações. Por mais que a quantidade de palavras-chave tenha apresentado uma diferença considerável entre os idiomas, como já discutido anteriormente, é interessante perceber que a quantidade total de candidatos a termo da moda não foi tão distante de um idioma para o outro, sendo Português: 22,0%; Francês: 24,0%; Inglês: 25,1%; Espanhol: 29,0%. Conclui-se que o tamanho do corpus de referência é importante no momento de coletar palavras-chave gerais, porém o corpus de estudo, por ser especializado e equilibrado, influencia mais na coleta dos candidatos a termo do que o tamanho do corpus de referência.

Gráfico 1 - Relação entre palavras-chave total, palavras-chave com frequência = ou + que 20 e total de candidatos a termo



Fonte: Da autora, 2023.

Ao analisar os termos coletados foi possível perceber que os termos com maior frequência no corpus são parte do léxico geral dos idiomas e não apenas do léxico especializado de moda, como é o caso dos termos saia, blusa e calça, por exemplo. Porém, apesar de serem palavras da língua geral, eles foram considerados como termo pois estão dentro da terminologia de moda e apresentam muitos *collocates* que são termos específicos de moda, por exemplo, no caso de saia evasê, saia lápis, calça capri, entre outros. Este trabalho buscou realizar um panorama geral de moda, com subáreas diferentes, por isso as palavras-chave com maior ocorrência no corpus acabam sendo parte da língua geral. Caso a escolha fosse por vídeos mais especializados, se os vídeos fossem apenas de tutoriais, por exemplo, o vocabulário também teria sido mais especializado. Também, devido à diferença de vocabulário entre os diferentes tipos de vídeos: tutoriais, tendências e história da moda, ao coletar os contextos de cada termo foi possível identificar que alguns termos aparecem apenas em certos tipos de vídeos, é o caso, por exemplo, de "costurar" que aparece apenas nos vídeos de tutoriais.

Alguns termos também foram considerados, apesar de não pertencerem apenas à terminologia de moda. Esses termos, que possuem pertinência pragmática e não temática, em sua maioria, fazem parte da subárea "cor" apresentado por Farias (2003), mas desconsiderada por Raimundo (2003). Sendo assim, infere-se que Farias (2003) também considerou termos de pertinência pragmática como parte importante da terminologia de moda, assim como este trabalho.

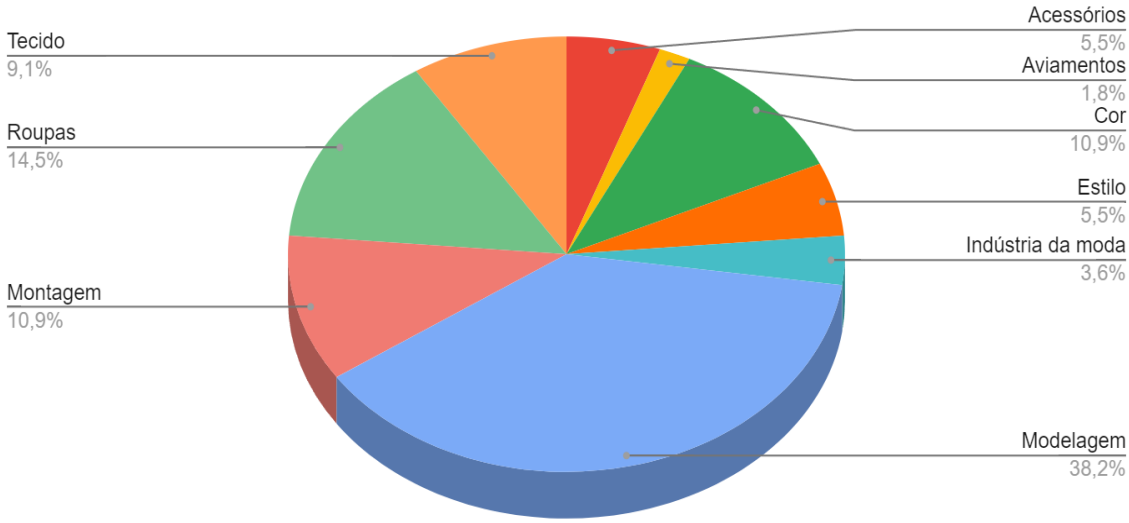
Termos como *dress*, *cut* e *top* apresentaram uma dificuldade, devido à polissemia dos termos, já que *dress*, por exemplo, pode ser tanto um verbo quanto um substantivo. Top aparece

tanto em francês quanto em inglês como sinônimo de blusa, no entanto, ele possui diferentes significados em ambos os idiomas. Em uma pesquisa futura seria interessante considerar essas diferenças na polissemia da terminologia da área, porém para esta pesquisa os termos foram definidos com apenas uma definição, que estava dentro da terminologia de moda, as demais foram desconsideradas.

Dentro da terminologia identificada nos idiomas português, francês e espanhol, alguns termos aparecem que vêm do inglês, como é o caso de *top* apresentado acima e de look, um neologismo que vêm do inglês, mas que ganhou novo significado dentro da área de moda, e que já está presente em glossários da área, definido pelo Dicionário da moda (TENDENZA DI MODA, 2019) como: "é a composição de peças, acessórios e estilo." e cujo equivalente em inglês foi "*outfit*" e não "*look*".

Na seção de metodologia, apresentou-se a quantidade de termos por subárea que apresentaram equivalências, no gráfico 2 abaixo é possível identificar a divisão desses termos em porcentagem. É possível notar que a subárea de "Modelagem" foi a mais representativa e possui uma quantidade consideravelmente maior de termos do que as outras subáreas. Isso pode se dar devido ao fato de que os termos de modelagem aparecem nos três tipos diferentes de vídeos que foram escolhidos para a pesquisa (história da moda, tendências e tutoriais), pois ele abrange ferramentas usadas no processo de modelagem e tipos de modelagem diferentes, como calça boca de sino e vestido tubinho. Já algumas das outras subáreas aparecem com mais frequência em tipos de vídeos específicos, é o caso da subárea "Montagem", por exemplo, os termos desta subárea aparecem em sua maioria nos vídeos de tutoriais, como citado anteriormente no caso do termo "costura", assim como os termos de "Aviamentos" e "Tecido".

Gráfico 2 - Quantidade de termos por subárea temática que possuem equivalências



Fonte: Da autora, 2023.

Como citado na seção de metodologia, os equivalentes encontrados foram divididos em aqueles que apareceram em até 2 idiomas diferentes, os que apareceram em até 3 idiomas, os que apareceram nos 4 idiomas propostos e os que não apresentaram equivalências. Este trabalho apresentou apenas um recorte dos dados coletados no corpus, por isso a quantidade de equivalências e termos não foi tão grande. No gráfico 3 abaixo é possível notar que quase 50% dos candidatos a termo identificados no português não apresentaram equivalências, isso pode acontecer devido a diferença de ocorrências entre os termos dentro do corpus, o que, ao realizar um recorte como o que foi feito aqui separa termos que poderiam ser equivalentes. Quanto mais próxima a frequência dos candidatos a termo no corpus, mais provável que elas fossem identificadas neste recorte, mas este não é o caso para todos os candidatos a termo. Neste caso, se os vídeos escolhidos tivessem sido mais específicos, de apenas uma subárea, por exemplo, mais equivalências teriam sido encontradas.

Gráfico 3 - Equivalentes que foram encontrados dos candidatos a termo, partindo do idioma português

Fonte: Da autora, 2023.

OBJ

4. Considerações Finais

Este trabalho buscou extrair e analisar candidatos a termo da área da moda, através da Linguística de corpus, usando o software Antconc e com a utilização de corpora orais. Como apresentado na seção de metodologia, a transcrição automática utilizada no corpus apresentou alguns problemas, por isso, cabe frisar que, com o avanço do processamento de linguagem natural e a melhora dos softwares de reconhecimento de voz, esses problemas vão sendo minimizados e os estudos que utilizam transcrições automáticas para analisar corpora orais ficarão cada vez mais confiáveis.

Destaca-se também que foi possível encontrar 56 termos em português com equivalências a partir de um recorte com apenas os 30 termos mais frequentes da pesquisa e seus *collocates*, sendo assim uma pesquisa que analise todas as 1606 palavras-chave identificados para o português e seus *collocates* encontrará uma quantidade muito maior de termos e equivalências. Além disso, é importante considerar que o corpus foi desenvolvido a partir da transcrição de 320 vídeos, porém a quantidade de material disponível no YouTube é bem superior a isso, e, quanto maior o corpus, mais candidatos a termo serão encontrados. A plataforma do YouTube é, na área de moda, a que apresenta uma grande quantidade de material. No entanto, ela não é a única plataforma que pode ser utilizada para esse tipo de pesquisa, os

podcasts, por exemplo, estão sendo difundidos e contém uma quantidade significativa que pode ser utilizado para pesquisa futura.

O objetivo deste trabalho não contemplava a criação do produto final do glossário, mas como pesquisa futura ficará a criação de um website chamado "Glomô - Glossário Multilíngue e Multimodal de Moda" usando Processamento de Linguagem Natural para que, a partir da planilha "Dossiê e equivalentes" o verbete para cada termo possa ser desenvolvido. A microestrutura contemplará as informações das fichas terminológicas: entrada principal, definição e exemplos. O glossário será multilíngue e multimodal. O código de programação será escrito em HTML, a partir do qual será possível realizar a conversão do texto de cada termo para áudio e para imagem, usando Inteligência Artificial. O processo será semiautomático, pois a Inteligência Artificial por estar no início, comete algumas falhas, sendo assim, será necessário um revisor para validar os dados gerados pelo programa.

5. REFERÊNCIAS

ABIT - ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL E DE CONFECÇÃO. **Perfil do setor**. Disponível em: <https://www.abit.org.br/cont/perfil-do-setor>. Acesso em: 4 set. 2023.

ALMEIDA, G. M. DE B.; CORREIA, M. Terminologia e corpus: relações, métodos e recurso. Em: **Avanços da Linguística de Corpus no Brasil**. São Paulo: Humanitas, 2008. p. 67–94.

ASSIS, M. B. C. G. D. **AVALIAÇÃO DA TERMINOLOGIA DE CORTE E COSTURA COM BASE EM CORPUS**. Trabalho de conclusão de curso—Brasília: Universidade de Brasília, 2020.

AUDACES. **Aprende los términos más importantes de la industria con este diccionario de moda**. Disponível em: <https://audaces.com/es/blog/diccionario-moda>. Acesso em: 10 nov. 2023.

BRUM-DE-PAULA, M. R.; ESPINAR, G. S. Coleta, transcrição e análise de produções orais. **Letras**, n. 21, 2002.

FARIAS, E. M. P. ASPECTOS SEMÂNTICOS DO LÉXICO DA MODA. **Rev. de Letras**, v. 1/2, n. 25, p. 73–79, dez. 2003.

FARIAS, E. M. P.; BEZERRA, T. M. F. Terminografia trilíngue. **Linguagem em Foco - Revista do Programa de Pós-Graduação em Linguística Aplicada da U ECE**, v. 1, n. 1, p. 51–51, 2009.

FIASCO, V. **The language of fashion as Language for Specific Purposes: collecting and analysing a trilingual comparable corpus drawn from the fashion magazines ELLE and VOGUE**. Doutorado—Itália: Università degli Studi Roma Tre, 2018.

GODOY, A. D. F. Apresentação do dicionário multilíngue de termos do setor feirístico: português, inglês, francês e italiano. **GTLex**, v. 4, n. 2, p. 317–334, jun. 2019.

KADER, C. C. C.; RICHTER, M. G. Linguística de corpus: possibilidades e avanços. **Instrumento: Revista de Estudo e Pesquisa em Educação Juiz de Fora**, v. 15, n. 1, p. 23, jun. 2013.

KAUR, J.; BUTTAR, P. A Systematic Review on Stopword Removal Algorithms. **International Journal on Future Revolution in Computer Science & Communication Engineering**, v. 4, n. 4, p. 207–210, 1 abr. 2018.

KRIEGER. Geração de glossários e dicionários especializados. Em: **Introdução à terminologia: teoria e prática**. São Paulo: Contexto, 2004. p. 127–144.

LEVI'S® BRASIL. **Dicionário do denim**. Disponível em: <https://www.levi.com.br/informacoes/guia-do-jeans/dicionario-do-denim>. Acesso em: 10 nov. 2023.

LOPES, L.; FINATTO, M. J. B.; CIULLA, A. Extração automática de candidatos a termos do Curso de Linguística Geral com apoio de recursos da Linguística de Corpus e do Processamento de Linguagem Natural. **Domínios de Lingu@gem**, v. 9, n. 2, p. 40–55, 18 dez. 2015.

MARTINS, A. F. R. DE O. F. et al. A call for a fashion pact: challenges and opportunities for circular economy in the brazilian fashion industry. **OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA**, v. 21, n. 7, p. 6168–6187, 5 jul. 2023.

NAGUMO, E.; TELES, L. F.; SILVA, L. DE A. A utilização de vídeos do Youtube como suporte ao processo de aprendizagem. **Revista Eletrônica de Educação**, v. 14, n. 3757008, p. 1–12, dez. 2020.

OLIVA, F. A.; BARBOSA, M. E. V. A. A IMPORT NCIA DOS IDIOMAS NAS RELAÇÕES DE COMÉRCIO EXTERIOR DO BRASIL. **Revista Alomorfia**, v. 7, n. 1, p. 641–650, 2 maio 2023.

OXFORD TEXT ARCHIVE. **British National Corpus, Baby edition**. Disponível em: <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2553>. Acesso em: 10 nov. 2023.

PAVEL, S.; NOLET, D. **MANUAL DE TERMINOLOGIA**. Tradução: Enilde Faulstich. Ministério de Obras Públicas e Serviços Governamentais do Canadá: Bureau de la traduction, 2002.

PRESTES, K. V. **Extração multilíngue de termos multipalavra em corpora comparáveis**. Mestrado—Porto Alegre: UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2015.

PROJETO AC/DC: CORPO CORPUS BRASILEIRO. **Acesso a corpos: corpo Corpus Brasileiro.** Disponível em: <https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS>. Acesso em: 10 nov. 2023.

RAIMUNDO, E. M. **UM ESTUDO TERMINOLÓGICO BILÍNGÜE (PORTUGUÊS-FRANCÊS) DO VOCABULÁRIO DA MODA: SUBÁREA VESTUÁRIO**. Mestre em Estudos da Linguagem—Londrina: Universidade Estadual de Londrina, 2003.

REAL ACADEMIA ESPAÑOLA. **Real Academia Española - CREA**. Disponível em: <https://corpus.rae.es/lfrecuencias.html>. Acesso em: 10 nov. 2023.

SARDINHA, T. B. **Linguística de Corpus**. Barueri, SP: Manole, 2004.

SARDINHA, T. B. A INFLUÊNCIA DO TAMANHO DO CORPUS DE REFERÊNCIA NA OBTENÇÃO DE PALAVRAS-CHAVE USANDO O PROGRAMA COMPUTACIONAL WORDSMITH TOOLS*. **the ESPecialist**, v. 26, n. 2, p. 183–204, 2005.

SIMÕES, J. **Glossário de Moda - Conheça os Termos da Moda. Blog Oscar**, 27 jun. 2023. Disponível em: <https://blog.oscarcalcados.com.br/glossario-de-moda/>. Acesso em: 10 nov. 2023

TABLOÏDE MODE. **Lexique de la mode | tabloïde mode**. Disponível em: <https://tabloidemode.media/lexique-de-la-mode/>. Acesso em: 10 nov. 2023.

TENDEZA DI MODA. **Dicionário da moda.** Disponível em: <https://tendenzadimoda.com.br/blog/dicionario-da-moda/>. Acesso em: 22 nov. 2023.

TIMANS, R.; WOUTERS, P.; HEILBRON, J. Mixed methods research: what it is and what it could be. **Theory and Society**, v. 48, n. 2, p. 193–216, 1 abr. 2019.

TRILOGY. **Denim terminology explained | Trilogy**. Disponível em: <https://www.trilogystores.co.uk/the-denim-guide/denim-terminology-explained/#>. Acesso em: 10 nov. 2023.

VAUGHAN, E.; O'KEEFFE, A. Corpus Analysis. Em: **The International Encyclopedia of Language and Social Interaction**. [s.l.] John Wiley & Sons, Ltd, 2015. p. 1–17.

WELCKER, H. A. **Dicionários: Uma pequena introdução à lexicografia**. 2. ed. [s.l.] Thesaurus, 2004. v. 1.

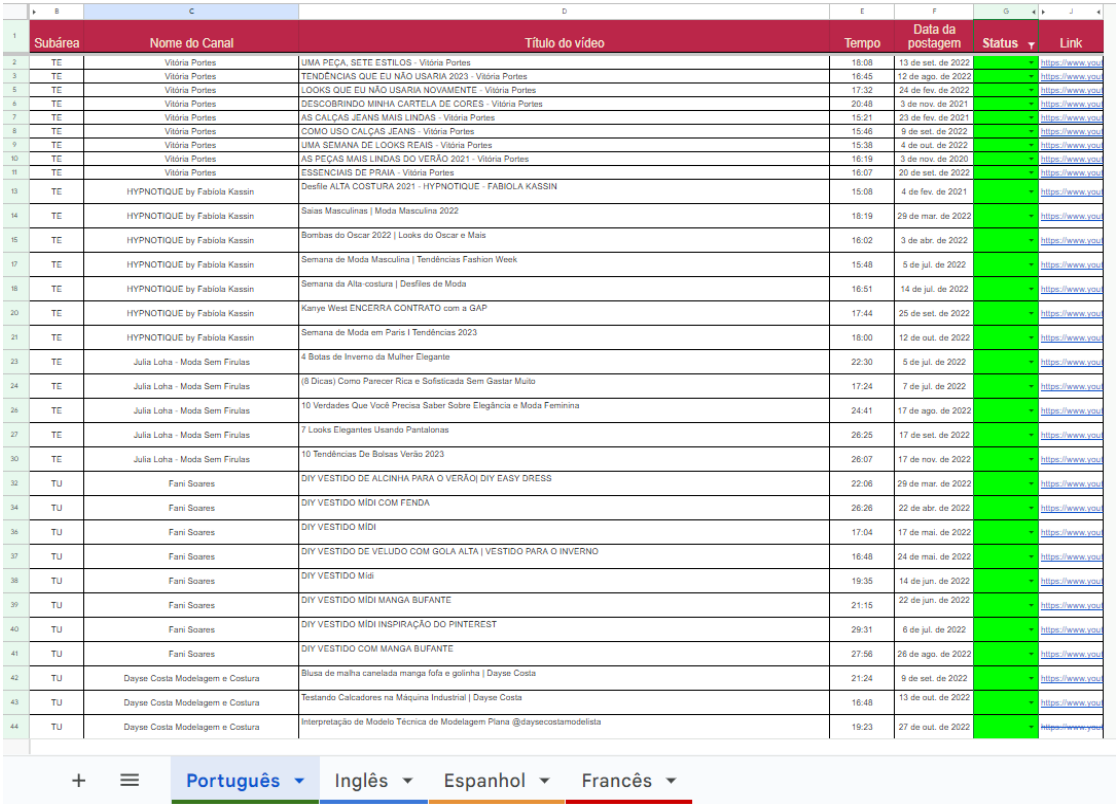
WORTSCHATZ LEIPZIG. **Download Corpora French**. Disponível em: <https://wortschatz.uni-leipzig.de/en/download/French>. Acesso em: 10 nov. 2023.

APÊNDICE

APÊNDICE A – Amostra da planilha criada para organizar as informações sobre cada youtuber: Primeiros youtubers dos idiomas português e inglês

Português		Nativo?	Formação	Profissão	Fonte	Inglês		Nativo?	Formação	Profissão	Fonte
Nome do canal	Nome da pessoa					Nome do canal	Nome da pessoa				
Vitória Portes	Jéssica Vitória Schoer Portes	Porto Alegre - RS	Ciências jurídicas e sociais	Designer de moda	https://www.youtube.com/watch?v=wZgWDpzT0rY&t=78s	Bernadette Banner	Bernadette Banner	Nova York - EUA	Bernadette graduated from New York University with a degree in theatrical production	Trabalhou na Broadway com design de moda	https://en.wikipedia.org/wiki/Bernadette_Banner https://www.coracaoboemio.com/bernadette-banner-uma-viajante-do-tempo/
HYPNOTIQUE by Fabíola Kassin	Fabíola Kassin	São Paulo	Gestão de turismo	Consultora de moda	https://www.linkedin.com/in/fabiolakassin/?originalSubdomain=br	With wendy	Wendy Liu	Toronto - Canadá	Business Administration and chemical engineering	Costureira Especialista de marketing	https://www.youtube.com/@withwendy/about https://www.linkedin.com/in/withwendy/?originalSubdomain=ca https://withwendy.com/about
Julia Loha - Moda Sem Firulas	Julia Loha	Rio de Janeiro	Design de moda	Designer de moda	https://www.youtube.com/@JuliaLohaModaSemFirulas/about https://www.linkedin.com/in/julia-loha-36222a157/?originalSubdomain=br	Rosery Apparel	Janelle	Tasmania - Australia		Designer de moda e costureira Trabalha em uma fábrica de tecidos	https://www.roseryapparel.com/about https://www.instagram.com/p/B3HJCoLn3ek/

APÊNDICE B - Planilha criada para a organização dos vídeos usados na criação dos corpora: exemplo dos primeiros vídeos em português

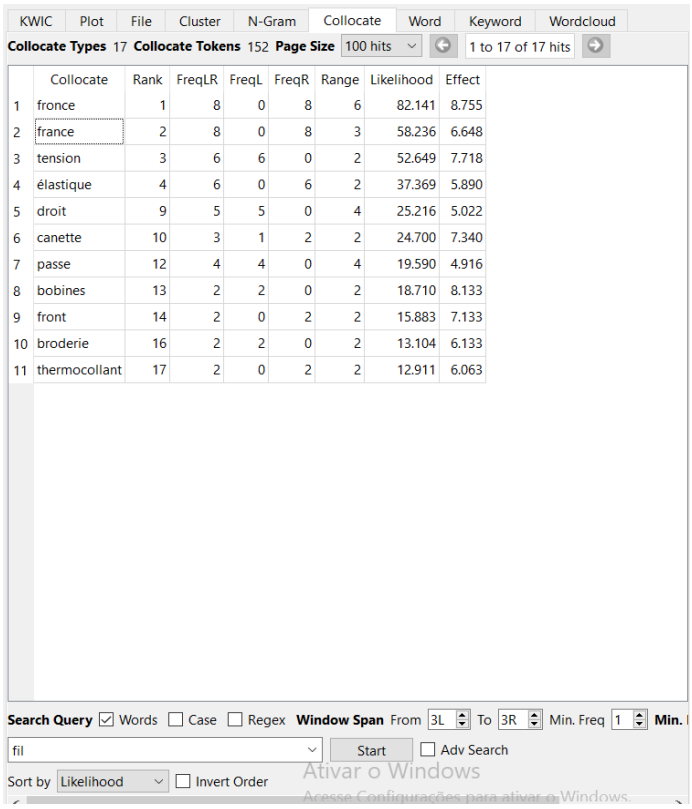


Subárea	Nome do Canal	Título do vídeo	Tempo	Data da postagem	Status	Link
TE	Vitória Portes	UMA PEÇA, SETE ESTILOS - Vitória Portes	18:08	13 de set. de 2022		https://www.you
TE	Vitória Portes	TENDÊNCIAS QUE EU NÃO USARIA 2023 - Vitória Portes	16:45	12 de ago. de 2022		https://www.you
TE	Vitória Portes	LOOKS QUE EU NÃO USARIA NOVAMENTE - Vitória Portes	17:32	24 de fev. de 2022		https://www.you
TE	Vitória Portes	DESCOBRINDO MINHA CARTELA DE CORES - Vitória Portes	20:48	3 de nov. de 2021		https://www.you
TE	Vitória Portes	AS CALÇAS JEANS MAIS LINDAS - Vitória Portes	15:21	23 de fev. de 2021		https://www.you
TE	Vitória Portes	COMO USO CALÇAS JEANS - Vitória Portes	15:46	9 de set. de 2022		https://www.you
TE	Vitória Portes	UMA SEMANA DE LOOKS REAIS - Vitória Portes	15:38	4 de out. de 2022		https://www.you
TE	Vitória Portes	AS PEÇAS MAIS LINDAS DO VERÃO 2021 - Vitória Portes	16:19	3 de nov. de 2020		https://www.you
TE	Vitória Portes	ESSENCIAIS DE PRAIA - Vitória Portes	16:07	20 de set. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Desfile ALTA COSTURA 2021 - HYPNOTIQUE - FABÍOLA KASSIN	15:08	4 de fev. de 2021		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Saias Masculinas | Moda Masculina 2022	18:19	29 de mar. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Bombas do Oscar 2022 | Looks do Oscar e Mais	16:02	3 de abr. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Semana de Moda Masculina | Tendências Fashion Week	15:48	5 de jul. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Semana da Alta-costura | Desfiles de Moda	16:51	14 de jul. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Kanye West ENCERRA CONTRATO com a GAP	17:44	25 de set. de 2022		https://www.you
TE	HYPNOTIQUE by Fabíola Kassin	Semana de Moda em Paris I Tendências 2023	18:00	12 de out. de 2022		https://www.you
TE	Julia Loha - Moda Sem Firulas	4 Botas de Inverno da Mulher Elegante	22:30	5 de jul. de 2022		https://www.you
TE	Julia Loha - Moda Sem Firulas	(8 Dicas) Como Parecer Rica e Sofisticada Sem Gastar Muito	17:24	7 de jul. de 2022		https://www.you
TE	Julia Loha - Moda Sem Firulas	10 Verdades Que Você Precisa Saber Sobre Elegância e Moda Feminina	24:41	17 de ago. de 2022		https://www.you
TE	Julia Loha - Moda Sem Firulas	7 Looks Elegantes Usando Pantalonas	26:25	17 de set. de 2022		https://www.you
TE	Julia Loha - Moda Sem Firulas	10 Tendências De Bolsas Verão 2023	26:07	17 de nov. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO DE ALCINHA PARA O VERÃO| DIY EASY DRESS	22:06	29 de mar. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO MÍDI COM FENDA	26:26	22 de abr. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO MÍDI	17:04	17 de mai. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO DE VELUDO COM GOLA ALTA | VESTIDO PARA O INVERNO	16:48	24 de mai. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO Midi	19:35	14 de jun. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO MÍDI MANGA BUFANTE	21:15	22 de jun. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO MÍDI INSPIRAÇÃO DO PINTEREST	29:31	6 de jul. de 2022		https://www.you
TU	Fani Soares	DIY VESTIDO COM MANGA BUFANTE	27:56	26 de ago. de 2022		https://www.you
TU	Dayse Costa Modelagem e Costura	Blusa de malha canelada manga fofa e golinha | Dayse Costa	21:24	9 de set. de 2022		https://www.you
TU	Dayse Costa Modelagem e Costura	Testando Calcadores na Máquina Industrial | Dayse Costa	16:48	13 de out. de 2022		https://www.you
TU	Dayse Costa Modelagem e Costura	Interpretação de Modelo Técnica de Modelagem Plana @daysecostamodelista	19:23	27 de out. de 2022		https://www.you

Português | Inglês | Espanhol | Francês

Legenda: As áreas dos canais foram abreviadas para TE (tendências), TU (tutoriais) e HM (história da moda); A coluna Status contém: concluída (verde) e não concluída (branco).

APÊNDICE C - Exemplo de erro de transcrição do termo “fil de fronce"



KWIC | Plot | File | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud

Collocate Types 17 **Collocate Tokens** 152 **Page Size** 100 hits 1 to 17 of 17 hits

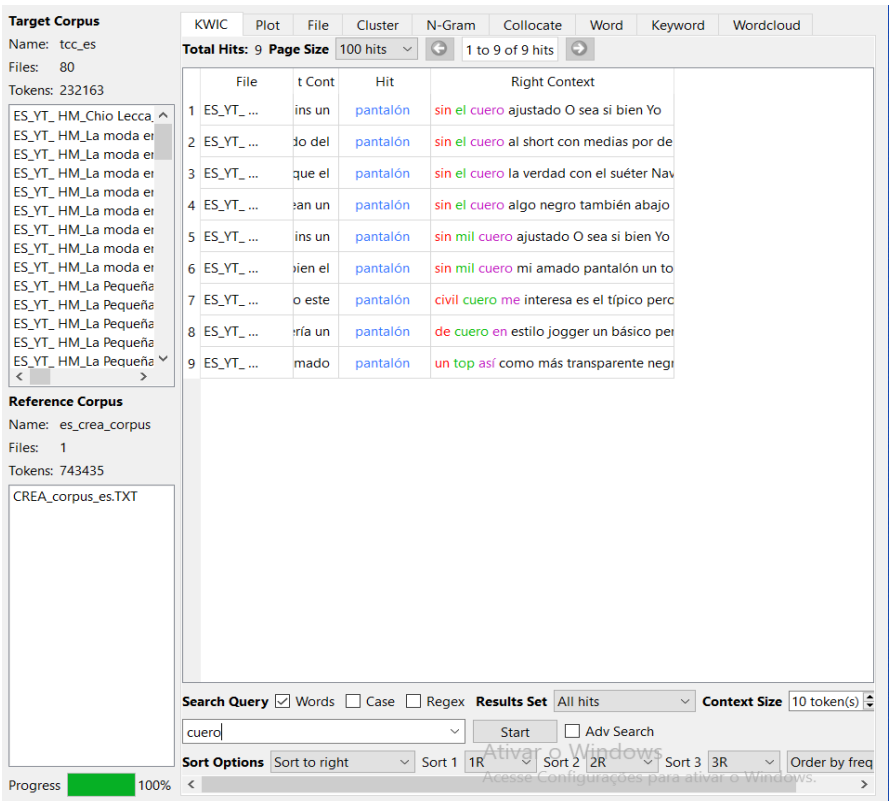
	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	fronce	1	8	0	8	6	82.141	8.755
2	france	2	8	0	8	3	58.236	6.648
3	tension	3	6	6	0	2	52.649	7.718
4	élastique	4	6	0	6	2	37.369	5.890
5	droit	9	5	5	0	4	25.216	5.022
6	canette	10	3	1	2	2	24.700	7.340
7	passe	12	4	4	0	4	19.590	4.916
8	bobines	13	2	2	0	2	18.710	8.133
9	front	14	2	0	2	2	15.883	7.133
10	broderie	16	2	2	0	2	13.104	6.133
11	thermocollant	17	2	0	2	2	12.911	6.063

Search Query Words Case Regex **Window Span** From 3L To 3R Min. Freq 1 **Min.**

fil Start Adv Search

Sort by Likelihood Invert Order

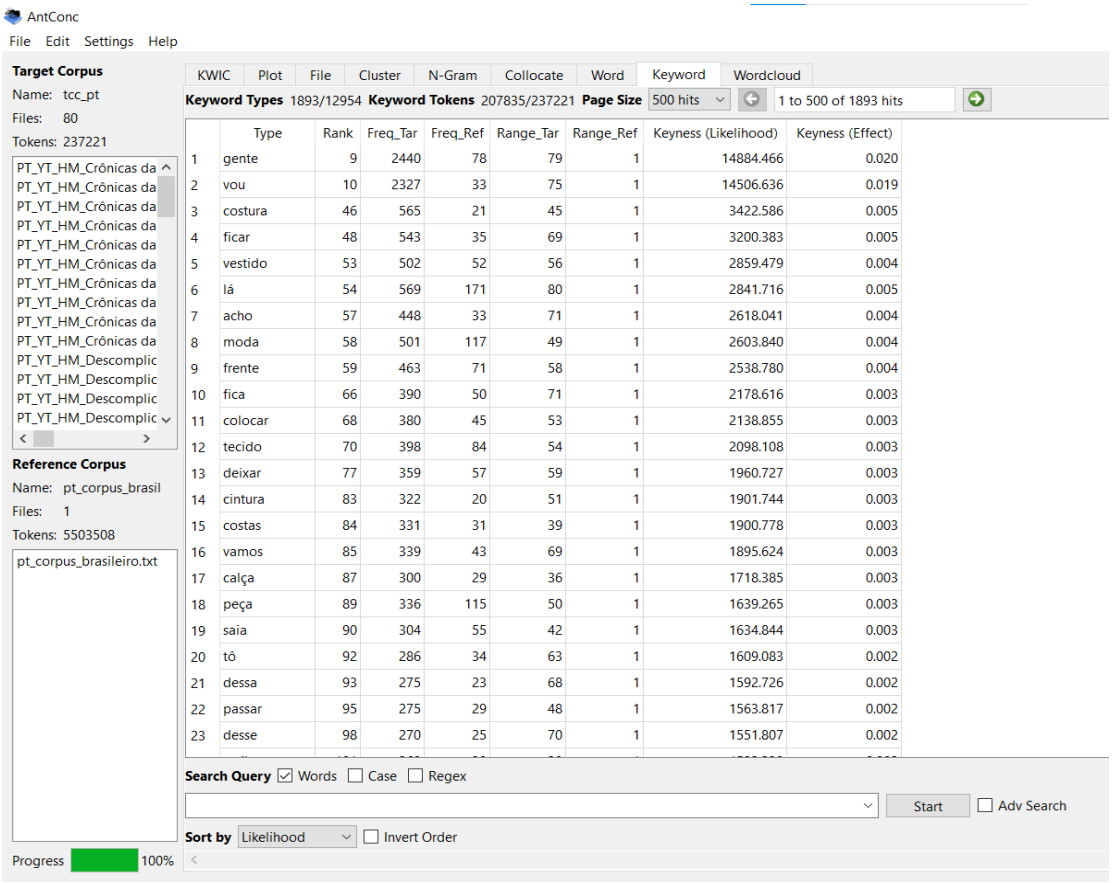
APÊNDICE D - Segundo exemplo de erro de transcrição: termo pantalón simil cuero



APÊNDICE E - Exemplo de nomeação dos arquivos do corpus em português

PT_YT_HM_Crônicas da Moda por Maria Landeiro_4 SÉRIES PARA ENTENDER A HISTÓRIA DA MODA Crônicas da Moda por Maria Landeiro_8 de jun. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_5 ÍCONES DE ESTILO DOS ANOS 70 Crônicas da Moda por Maria Landeiro_31 de ago. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_5 ÍCONES DE ESTILO DOS ANOS 80 Crônicas da Moda por Maria Landeiro_19 de out. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_A HISTÓRIA DA BURBERRY CRÔNICAS DA MODA POR MARIA LANDEIRO_26 de out. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_A HISTÓRIA DA CAROLINA HERRERA Crônicas da Moda por Maria Landeiro_30 de nov. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_A HISTÓRIA DA TIFFANY & CO. Crônicas da Moda por Maria Landeiro_11 de mai. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_A MODA DIPLOMÁTICA DA RAINHA ELIZABETH II Crônicas da Moda por Maria Landeiro_22 de set. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_ALEXANDER MCQUEEN - A HISTÓRIA DO GENIO Crônicas da Moda por Maria Landeiro_28 de set. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_BLONDE - TUDO SOBRE OS FIGURINOS Crônicas da Moda por Maria Landeiro_5 de out. de 2022
PT_YT_HM_Crônicas da Moda por Maria Landeiro_DOLCE & GABBANA - HISTÓRIA E POLÊMICAS CRÔNICAS DA MODA POR MARIA LANDEIRO_7 de set. de 2022
PT_YT_HM_Descomplicando a moda_A GRANDE LOJA DE DEPARTAMENTOS DE PARIS "LE BON MARCHÉ" DESCOMPLICANDO A MODA_17 de set. de 2021
PT_YT_HM_Descomplicando a moda_COMO A COR BRANCA VIROU MODA DESCOMPLICANDO A MODA_9 de set. de 2021
PT_YT_HM_Descomplicando a moda_LA SAMARITAINE A LOJA QUERIDINHA DE EMILY IN PARIS DESCOMPLICANDO A MODA_3 de jan. de 2022
PT_YT_HM_Descomplicando a moda_O STREETWEAR JAPONÊS SEMANA ESPECIAL JAPÃO DESCOMPLICANDO A MODA_13 de ago. de 2021
PT_YT_HM_Descomplicando a moda_OS MELHORES LIVROS DE MODA QUE COMPREI DESCOMPLICANDO A MODA_10 de jan. de 2022
PT_YT_HM_Descomplicando a moda_PERFUME CHANEL Nº 5 DESCOMPLICANDO A MODA_27 de jan. de 2022
PT_YT_HM_Descomplicando a moda_ROCHAS - O COSTUREIRO DAS MULHERES DESCOMPLICANDO A MODA_30 de set. de 2021
PT_YT_HM_Descomplicando a moda_UNIFORMES E A MODA SEMANA ESPECIAL JAPÃO DESCOMPLICANDO A MODA_10 de ago. de 2021
PT_YT_HM_História da Moda_ A História da FAMÍLIA GUCCI - Fundação, Legado e ASSASSINATO_28 de nov. de 2021
PT_YT_HM_História da Moda_A História das BOLSAS nos ANOS 2000 Jeans, Betty Boop, Tiracolo e outras - História da Moda_26 de out. de 2022
PT_YT_HM_História da Moda_A MODA de ELVIS PRESLEY - Como o Rei do Rock se vestia_13 de jul. de 2022
PT_YT_HM_História da Moda_A MODA nos ANOS 1920 - Melindrosas, Art Déco, Silhueta e mais..._11 de mai. de 2022
PT_YT_HM_História da Moda_A MODA nos ANOS 1980 - Características Principais_3 de abr. de 2022
PT_YT_HM_História da Moda_LEGALMENTE LOIRA - O Figurino de ELLE WOODS - Moda ANOS 2000_10 de nov. de 2021
PT_YT_HM_História da Moda_Moda na ERA EDUARDIANA (1900-1910) Como as mulheres se vestiam no início do século 20_6 de jan. de 2022
PT_YT_HM_História da Moda_O FIGURINO de O CRAVO e A ROSA Foi Uma Boa Representação dos ANOS 20 Análise_14 de nov. de 2022
PT_YT_HM_História da Moda_Tendências que SÓ FORAM bonitas na Moda dos ANOS 80_21 de ago. de 2022
PT_YT_HM_Lilian Pacce_DECIFRANDO OS LOOKS DA POSSE DE JOE BIDEN DE LADY GAGA A KAMALA HARRIS LILIAN PACCE_24 de jan. de 2021
PT_YT_HM_Lilian Pacce_GRAMMYS 2022 Decifrando os looks!_6 de abr. de 2022
PT_YT_HM_Lilian Pacce_O NOVO LOOK DA TURMA DE SEX AND THE CITY, AGORA EM AND JUST LIKE THAT! LILIAN PACCE #satc_21 de nov. de 2021
PT_YT_HM_Lilian Pacce_THE CROWN ensina lições de estilo do mar de lama ao mar de pérolas! LILIAN PACCE_28 de nov. de 2020

APÊNDICE F - Lista de palavras-chave em português gerada pelo software AntConc

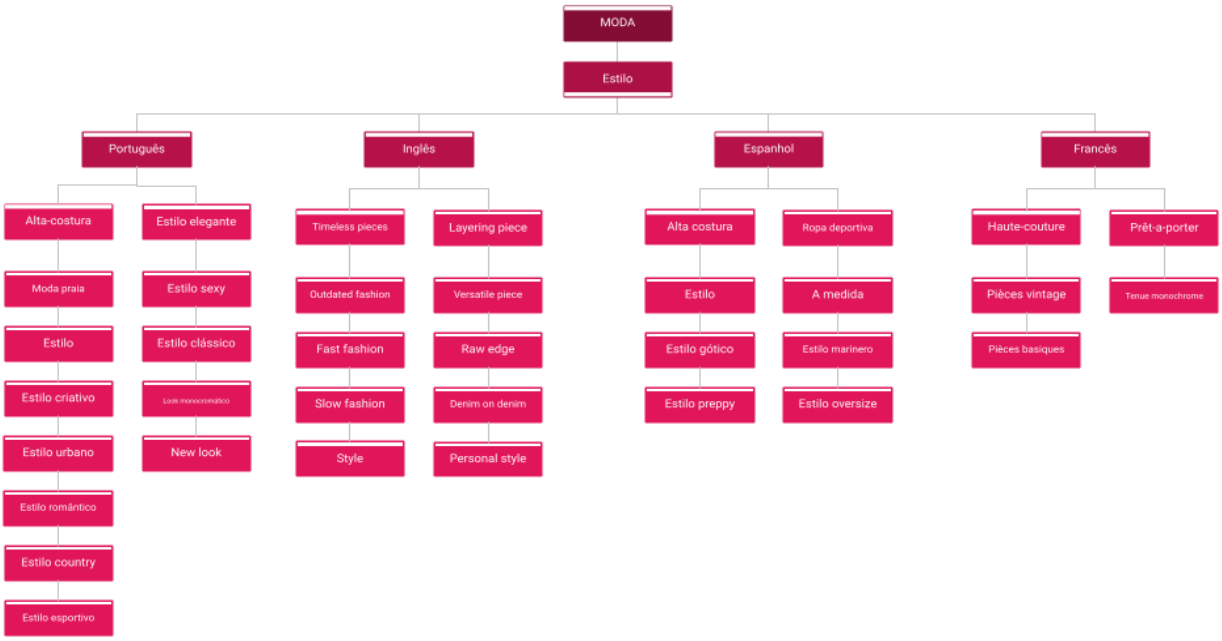


	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	gente	9	2440	78	79	1	14884.466	0.020
2	vou	10	2327	33	75	1	14506.636	0.019
3	costura	46	565	21	45	1	3422.586	0.005
4	ficar	48	543	35	69	1	3200.383	0.005
5	vestido	53	502	52	56	1	2859.479	0.004
6	lá	54	569	171	80	1	2841.716	0.005
7	acho	57	448	33	71	1	2618.041	0.004
8	moda	58	501	117	49	1	2603.840	0.004
9	frente	59	463	71	58	1	2538.780	0.004
10	fica	66	390	50	71	1	2178.616	0.003
11	colocar	68	380	45	53	1	2138.855	0.003
12	tecido	70	398	84	54	1	2098.108	0.003
13	deixar	77	359	57	59	1	1960.727	0.003
14	cintura	83	322	20	51	1	1901.744	0.003
15	costas	84	331	31	39	1	1900.778	0.003
16	vamos	85	339	43	69	1	1895.624	0.003
17	calça	87	300	29	36	1	1718.385	0.003
18	peça	89	336	115	50	1	1639.265	0.003
19	saia	90	304	55	42	1	1634.844	0.003
20	tô	92	286	34	63	1	1609.083	0.002
21	dessa	93	275	23	68	1	1592.726	0.002
22	passar	95	275	29	48	1	1563.817	0.002
23	desse	98	270	25	70	1	1551.807	0.002

APÊNDICE G - Planilha “Dossiê e equivalentes” criada para armazenar os dados sobre os termos

Pertinência	Subárea temática	Palavras-chave	Colocates	Frequência no corpus	Definição	Fonte	Contexto	Fonte
Temática	Montagem	costura		565	1. ato, modo ou efeito de coser, de unir por meio de pontos dados com agulha enfiada em linha 2. a linha que une as partes costuradas 3. arte de coser; atividade de confeção de vestuário; ofício de costureira, modista ou alfaiate	https://www.infopedia.pt/dicionarios/lingua-portuguesa/costura	1. Eu vou cortar ele aqui com 1 centímetro de sobra de costura em toda volta e agora pessoal, dobrei meu tecido aqui 2. Onde acaba a costura do zíper, eu vou continuar com outra costura até a mesma altura da costura da outra lateral	PT_YT_TU_Viviane Prates Modelagem e Costura_CORTE E COSTURA VESTIDO LONGO ESTILOSO COM LASTEX E BARRADO | VIVIANE PRATES_15 de set. de 2022.txt PT_YT_TU_Fani Soares_DIY VESTIDO Midi_14 de jun. de 2022.txt
Temática	Montagem		margem de costura	46	É a margem deixada no molde para a execução da costura. Toda peça precisa de margem de costura, sem ela, ao costurar, ocorre diminuição nas medidas, o que altera em todo o processo final da peça.	https://patriadacostura.blogspot.com/2014/07/meu-dicionario-da-costura.html	1. Eu deixei a lateral com 3 cm de margem de costura, porque vai ter um zíper e porque esse tecido não tem elastano 2. Eu vou cortar de novo desse jeito aqui na dobra do tecido, vou deixar 1 cm de margem de costura em toda a volta	PT_YT_TU_Fani Soares_DIY VESTIDO DE VELUDO COM GOLA ALTA VESTIDO PARA O INVERNO_24 de mai. de 2022.txt PT_YT_TU_Viviane Prates Modelagem e Costura_CORTE E COSTURA VESTIDO LONGO ESTILOSO COM LASTEX E BARRADO | VIVIANE PRATES_15 de set. de 2022.txt
Temática	Estilo		alta-costura	55	são peças reconhecidas pela comissão sediada no Ministério da Indústria Francesa como roupas de luxo.	https://blog.oscarcalcados.com.br/glossario-de-moda/	1. As revistas femininas tinham um papel crucial ao ditar as tendências e mostrar com exclusividade os lançamentos da alta-costura 2. As casas de alta costura viveram talvez a sua grande era de ouro nos anos 20	PT_YT_HM_História da Moda_A MODA nos ANOS 1920 - Melindrosas, Art Déco, Silhueta e mais..._11 de mai. de 2022.txt PT_YT_HM_História da Moda_O FIGURINO de O CRAVO e A ROSA Foi Uma Boa Representação dos ANOS 20 Análise_14 de nov. de 2022.txt
Temática	Montagem		costura reta	38	A costura reta é aquela que vai ser executada a partir de um ponto fixo formado por duas linhas, a superior e a inferior, e ao longo da costura, elas serão entrelaçadas umas às outras. A linha superior é aquela que passa pela agulha, já a inferior é a que se encontra na bobina	https://blog.maximustecidos.com.br/tudo-sobre-a-costura-reta/#:~:text=A%20costura%20reta%20%C3%A9%20aquela,ser%C3%A3o%20entrela%C3%A7adas%20umas%20%C3%A0s%20outras.	1. Vou passar agora uma costura reta com ponto normal depois de fazer o acabamento no zigue-zague 2. Vou fechar aqui os ombros, as laterais e o entrepernas aqui do short com a costura reta também	PT_YT_TU_Viviane Prates Modelagem e Costura_CORTE E COSTURA VESTIDO LONGO ESTILOSO COM LASTEX E BARRADO | VIVIANE PRATES_15 de set. de 2022.txt PT_YT_TU_Dayse Costa Modelagem e Costura_Macaquinho Abotoamento frontal Molde e Costura @daysecostamodelista_8 de dez. de 2022.txt
Temática	Roupas	vestido		502	Vestimenta feminina inteiriça que cobre o corpo inteiro, formada de saia e blusa em geral feitas de uma só peça de pano	https://www.dicio.com.br/vestido/	1. Vamos passar uma costura desde aqui da manga até lá na barra do vestido, aqui vamos passar o lastex 2. Um ícone fashion dos anos 80 foi a princesa Diana, seu vestido usado no casamento com Príncipe Charles no ano de 1981 inspirou milhares de noivas	PT_YT_TU_Viviane Prates Modelagem e Costura_CORTE E COSTURA VESTIDO LONGO ESTILOSO COM LASTEX E BARRADO | VIVIANE PRATES_15 de set. de 2022.txt PT_YT_HM_História da Moda_A MODA nos ANOS 1980 - Características Principais_3 de abr. de 2022.txt

APÊNDICE H - Organograma da subárea estilo contendo os candidatos a termo nos 4 idiomas



APÊNDICE I - Recorte da planilha "Dossiê e equivalentes" que apresenta os equivalentes encontrados nos idiomas

Entrada Principal		Equivalentes					
Português	Frequência	Inglês	Frequência	Espanhol	Frequência	Francês	Frequência
alta-costura	56			alta costura	111	haute-couture	44
avesso da peça	5	wrong side	14				
baby look	4						
base da blusa	8						
base da manga	6						
blusa	**253**	**top**	**437**	**blusa**	**184**	**top**	**167**
blusa de paetê	3	sequin pieces	3				
blusa listrada	4						
blusa ombro a ombro	11						
bolsa	177	bag	148			sac	189
bolsa tiracolo	5					sac à main	17
bolsa baguete	14					sac baguette	4
branco	187	white	200			blanc	147
cabeça da manga	9						
calça	300			pantalón	275	pantalon	214
calça boca de sino	3					pantalon flair [flare]	6
calça capri	2			pantalón capri	3		
calça cintura baixa	7						
calça jeans	61	jeans	151	pantalón vaquero	5		
calça legging	5						
calça mom jeans	2						
calça saruel	2						
calça skinny	11	skinny jeans	19				

Legenda: As linhas foram pintadas de acordo com a quantidade de equivalentes, em preto para os equivalentes nos 4 idiomas, em rosa para os equivalentes em até 3 idiomas e em azul para os equivalentes em até 2 idiomas.