Os estágios da vida humana em representações sociais: uma perspectiva da Linguística de *Corpus*

DOI: http://dx.doi.org/10.21165/el.v50i1.3077

Barbara Soares da Silva[1]

Resumo

O artigo que aqui se apresenta foi motivado pela necessidade de compreender as representações do ser humano nas diversas fases da vida. Estas foram operacionalizadas por meio dos seguintes itens lexicais: *man*, *woman*, *adolescent*, *adolescence*, *adult*, *boy*, *girl, child, kid, teen* e *teenager* (bem como em suas formas plurais). Os dados da análise consistiram nas publicações disponibilizadas pelo Google Books no período de 1800 a 2008. Para tanto, foram realizadas diversas análises de sequências de palavras adjacentes formadas pela base de dados do Google Books *Ngrams*. A pesquisa fundamenta-se na Linguística de *Corpus*, por meio da qual foi possível verificar os padrões de uso dessas palavras bem como a variação do uso desses itens. O estudo compreendeu análise quantitativa, primeiramente, e qualitativa, posteriormente, por meio da interpretação da temática apontada pelo léxico, pela leitura e análise de textos disponibilizados pela base do Google Books. A partir da análise dos padrões e da variação temporal de uso, foram apontadas as representações emergentes de cada item investigado. Com base nessa análise, foi possível detectar presença das representações, sendo possível verificar como o ser humano tem sido representado pela linguagem (em inglês) nos últimos três séculos. Além dos resultados obtidos, a presente pesquisa sugere o poder do estudo da análise histórica baseada em grandes quantidades de dados textuais.

Palavras-chave: Linguística de *Corpus*; Google Books; representação social; Análise de Sentimento.

1 Pontifícia Universidade Católica (PUC), São Paulo, São Paulo, Brasil; casadatraducao@gmail.com; http://orcid.org/0000-0002-7067-5594

The stages of human life in social representations: a Corpus Linguistics perspective

Abstract

The article presented here has been motivated by the need to understand representations of the human being in the different stages of life. These were operationalized through the following lexical items: man, woman, adolescent, adolescence, adult, boy, girl, child, kid, teen, and teenager (as well as in their plural forms). The analysis data consisted of the publications enabled by Google Books from 1800 to 2008. Therefore, several analyzes of sequences of adjacent words created by the Google Books Ngrams database were carried out. The research is based on Corpus Linguistics, by which it was possible to verify the usage patterns of these words as well as the variation in the use of such items. The study comprised quantitative analysis, first, and qualitative, later, through the interpretation of the theme pointed out by the lexicon, by reading and analyzing texts provided by the Google Books database. From the analysis of the patterns and the *temporal* variation of use, the emerging representations of each item investigated were highlighted. Based on such analysis, it was possible to detect the presence of the representations, being possible to verify how the human being has been represented by language (in the English language) in the last three centuries. In addition to the results obtained, this research highlights the power of the study on historical analysis based on large amounts of textual data.

Keywords: Corpus Linguistics; Google Books; social representation; sentiment analysis.

Considerações iniciais

A pesquisa desenvolvida por ocasião da realização do Doutorado em Linguística na Pontifícia Universidade Católica de São Paulo, no Programa de Linguística Aplicada e Estudos da Linguagem, identificou e estudou as representações sociais linguísticas do Ser Humano ao longo de 21 décadas em língua inglesa por meio do uso do Google Books *Ngram*. O presente estudo está fundamentado no âmbito da Linguística de *Corpus* (LC) e, portanto, na Linguística Aplicada (LA).

Sendo assim, para dar início, têm-se que a Linguística de *Corpus* é uma das vertentes da Linguística Aplicada (LA) e centra-se na resolução de problemas de uso da linguagem, e por esta razão é uma ciência social (ALVAREZ; SILVA, 2007).

A relevância do relato do presente estudo, concluído em meados de 2019, reside no fato de que ainda não há uma análise com base na Linguística Aplicada deste teor na literatura, tampouco foram encontradas publicações na área voltando-se em especial ao aspecto aqui explorado.

Este artigo relata um estudo em que o objetivo consistiu em investigar padrões linguísticos frequentes de palavras voltadas ao Ser Humano em inglês, sendo elas: *man* (homem), *woman* (mulher), *adolescent* (adolescente), *adolescence* (adolescência), *adult* (adulto), *boy* (menino), *girl* (menina), *child* (criança), *kid* (criança), *teen* (adolescente) e *teenager* (adolescente), bem como suas formas plurais, a partir de dados do Google Books que cobrem o período de 1800 a 2008. Pretendeu-se também identificar os padrões das palavras e verificar se há mudanças em relação aos padrões no período das vinte e uma décadas do estudo. Dessa forma, o objetivo consistiu em responder as seguintes perguntas de pesquisa: *1) quais representações podem ser identificadas em relação aos termos pesquisados? 2) há diferença entre as representações dos termos masculinos e femininos? E entre os infantis, adolescentes e adultos? 3) há diferença entre os termos em relação à valoração (carga positiva e negativa)?* Neste artigo, serão respondidas as perguntas 1 e 3.

De acordo com Stubbs (1996), os padrões de uso dos itens lexicais podem sinalizar a representação que esses itens assumem na sociedade: as maneiras recorrentes de falar não determinam o pensamento. Estes oferecem representações convencionais ou de pessoas e acontecimentos por meio do filtro e da cristalização de ideias além de prover significados pré-fabricados nos quais podem ser facilmente captados e veiculados. O autor buscou quais foram as colocações mais frequentes destas palavras, ou seja, os padrões linguísticos formados pela presença de duas palavras próximas uma à outra (geralmente separadas por até quatro outras palavras ou colocados), por exemplo, "British Empire". Por colocados ou *collocations*, se compreende como palavras que naturalmente estão em conjunto e, assim, cristalizam um significado.

Há demais estudos disponíveis em Linguística de *Corpus* no campo das representações que também estabelecem a mesma relação entre o uso frequente de determinados padrões linguísticos e a presença de representações, como Baker (2014), Baker e Potts (2013) e Baker e Ellece (2011). Embora haja estudos com base em *corpora* sobre representações, não há pesquisas que investiguem as representações em torno de itens lexicais relativos ao ser humano e aos seus estágios da vida. Uma vez que não há precedentes de estudos dentro da Linguística de *Corpus* dedicados à investigação da representação social em relação ao ser humano, o artigo que relata tal pesquisa pretende preencher esta lacuna e descrever como ela se dá.

Fundamentação teórica

Conforme mencionado nas Considerações iniciais deste artigo, a seguir apresentamos o detalhamento de alguns conceitos. A iniciar pela Linguística de *Corpus*, que representa hoje a força-motriz para engendrar os avanços de método e teoria empregados neste estudo, têm-se nas palavras de Berber Sardinha (2004, p. 3),

> A conceituação inicial da Linguística de *Corpus*, fundamentalmente sendo: a área da linguística que se ocupa da coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística.

Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. Similarmente, são muitos os autores que definem a Linguística de *Corpus*. Alguns exemplos de vozes-chave são: Sinclair (1991), Stubbs (1993), McEnery e Wilson (1996), Biber, Conrad e Reppen (1998), Kennedy (1998), Hunston (2002) e Berber Sardinha (2004). A Linguística de *Corpus* pode ser expressamente entendida como o estudo da linguagem baseado em exemplos da língua usada na "vida real" (McENERY; WILSON, 1996).

Assumidamente no tangente à importância mais que atual desta área da Linguística Aplicada, Biber (2010) coloca que a Linguística de *Corpus* é muito mais do que uma abordagem metodológica; as inovações metodológicas possibilitaram que pesquisadores façam perguntas de pesquisa até então inéditas. Biber (2010) sugere ainda que a Linguística de *Corpus* oferece um suporte robusto para a visão empírica de métodos quantitativos e contribui cada vez mais para a noção de que a principal contribuição da Linguística de *Corpus* está, de fato, em ser capaz de documentar o uso de diversas características linguísticas, incluindo a variação de seu uso. Há muitas definições de *corpus*, mas todas centram-se na ideia de que um *corpus* é uma coletânea de textos em formato eletrônico produzidos em condições reais de uso. Segundo Sinclair (2005, p. 4, tradução minha[2]), *corpus* na perspectiva da Linguística de *Corpus* é: "Conjunto de partes de uma linguagem em texto, em formato eletrônico, selecionado de acordo com critérios externos a serem representados; uma linguagem ou variedade linguística como fonte de dados para pesquisa linguística".

Para Trask (2004, p. 64), *corpus* é "um conjunto de textos escritos ou falados numa língua, disponível para análise". Para Galisson e Coste (1983), *corpus* é um conjunto finito de enunciados tomados como objeto de análise reunidos para servir de base à descrição. Já para Dubois *et al.* (1993), *corpus* é um conjunto de enunciados a partir do qual se pode descrever a gramática de uma língua. Na concepção de Ducrot e Todorov (2001, p. 32), *corpus* é um "conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua em determinada época".

Na Linguística de *Corpus*, tem-se como base o estudo de padrões de linguagem. Assim, segundo Berber Sardinha (2004, p. 39),

2 No original: "A corpus is a collection of texts. More specifically, in the words of Sinclair, it is "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language".

> Os padrões podem ser definidos como a associação entre palavras e estruturas. A importância da padronização reside no fato de que certos significados emanam dessa associação e, portanto, o estudo da padronização envolve o estudo dos significados da língua em uso.

Neste estudo, foi efetuada a análise dos padrões de linguagem de um conjunto de palavras, em inglês, referentes ao ser humano e, a partir desses padrões, tentamos verificar as representações associadas às diferentes formas de se referir ao ser humano ao longo do tempo. Um padrão pode ser identificado se uma combinação de palavras ocorre com relativa frequência e se há um significado associado.

Análise de Sentimento (*Sentiment Analysis*) para o presente estudo

Na pesquisa narrada neste artigo, propõe-se o conceito de "valoração" para mensurar a prosódia semântica. Normalmente, a prosódia semântica é analisada qualitativamente, por meio da verificação dos usos da palavra. Geralmente essa análise é conduzida por meio do exame de itens individuais e não tem como objetivo atribuir um índice numérico aos itens analisados. Por outro lado, é necessário analisar centenas de itens e, portanto, foi averiguada uma maneira de empreender a análise de modo automático, quantitativamente. Para o presente estudo, baseou-se na Análise de Sentimento (*Sentiment Analysis*), por meio da qual foi possível atribuir um valor numérico a cada um dos itens investigados. Assim, valoração, nesta pesquisa, significa uma medida do valor, em termos de positividade ou negatividade, que uma palavra possui tendo em vista seu uso. Por exemplo, a palavra inglesa "*wretched*" (miserável, amaldiçoado, etc.) possui valoração geralmente negativa com um índice de valoração de -3.43, segundo Hamilton, Leskovec e Jurafsky (2016). O índice, sendo negativo, significa que o item de valoração negativa apresenta um sentimento emocional negativo por parte do colocado analisado – da mesma maneira que, na classificação geral, este tipo de sentimento pode representar também um processo ou um estado, como na menção a seguir:

> Os campos foram sendo gradualmente cobertos por minas, fundições, fábricas e oficinas e fileiras de casebres miseráveis para os homens, mulheres e crianças que trabalhavam neles: uma grande conurbação industrial não planejada que era sombria. Durante o dia e assustadora à noite' de David Lodge, na obra Changing Places. (LODGE, 1975) (tradução minha[3]).

3 No original: "The fields were gradually covered with pitheads, foundries, factories and workshops and rows of wretched hovels for the men, women and children who worked in them: a sprawling, unplanned, industrial conurbation that was gloomy by day, fearsome by night' de David Lodge, na obra Changing Places".

Dessa forma, sendo o autor mencionado pioneiro no tema, tem-se que a Análise de Sentimento é uma linha de pesquisa em Linguística Computacional e Processamento de Linguagem Natural que tem como objetivo mensurar a valoração dos "sentimentos" expressos na linguagem. Uma das aplicações principais dessa linha é na análise da valoração de textos de redes sociais, em tempo real, ou seja, à medida em que os textos postados nas redes vão sendo produzidos. Por exemplo, uma empresa ou partido político pode ter interesse em saber como o público está reagindo a algum produto, imagem corporativa ou campanha política nas redes sociais, em termos de se o público está reagindo de modo favorável ou não a essas questões. Por meio de algoritmos computacionais, o "sentimento" expresso pelo público nas postagens é mensurado e o resultado é mostrado em termos de quão positiva ou negativa está sendo a reação.

Em Hamilton, Leskovec e Jurafsky (2016), de acordo com o trecho originalmente em Língua Inglesa, o algoritmo desenvolvido pelos autores foi o SENTPROP, que retorna um índice de valoração para cada palavra estudada. A importância do estudo de Hamilton, Leskovec e Jurafsky (2016) para esta pesquisa é o fato de os índices de valoração da pesquisa terem sido disponibilizados na forma de dicionários (listagens) prontos.

Os autores empregaram o *corpus* histórico COHA (*Corpus of Historical American English*), com cerca de 400 milhões de palavras para mensurar a polaridade positiva ou negativa do léxico de língua inglesa. Assim, os dicionários de polaridade vieram indexados pela época em que o item foi usado. Para os propósitos da pesquisa descrita aqui neste artigo, o fato de existir um dicionário de análise de sentimento discriminado por período de tempo é importante, uma vez que os dados desta pesquisa também são discriminados temporalmente.

Segundo Moscovici (1988), as representações sociais dizem respeito aos conteúdos de raciocínio rotineiro e o armazenamento de ideias que oferecem coerência às crenças religiosas, ideias políticas e conexões que nós criamos espontaneamente da mesma forma que o ar que respiramos. Torna-se possível, assim, classificar pessoas e objetos para comparar e explicar comportamentos e objetivar isto como parte integrante de nosso contexto social. Enquanto as representações são com frequência localizadas nas mentes dos homens e das mulheres, estas podem também com frequência ser encontradas no mundo e, conforme são, ser examinadas separadamente. Segundo Stubbs (1996), as representações que circulam na língua podem ser verificadas por meio da análise de *corpora*. O autor apresenta um estudo baseado em *corpus* que teve como objetivo identificar as representações de itens como "British" e "English" em um *corpus* de notícias de jornal em inglês dos anos de 1990. Os resultados mostraram que "British" às vezes possui uma representação neutra, como é o caso do adjetivo associado a nomes de instituições e corporações, tais como, Airways e Telecom. Assim, o uso da Linguística de *Corpus*, juntamente com a Análise de Sentimento (*Sentiment Analysis*) e os conceitos das Representações Sociais perfizeram a fundamentação teórica para o desenvolvimento da pesquisa descrita neste artigo. A seguir a descrição dos passos descritos na metodologia desta.

Procedimentos metodológicos

Corpus de pesquisa

Os dados empregados na pesquisa aqui relatada trataram de listagens de ocorrências de bigramas encontrados nas publicações em língua inglesa indexadas pelo Google Books. O Google Books é uma coletânea de milhões de publicações digitalizadas pela empresa Google a partir dos acervos de bibliotecas ao redor do mundo. Já os bigramas são sequências de duas palavras colocadas lado a lado em um texto. Por exemplo: "Brazilian poet", "young women" e "American men". Os bigramas são disponibilizados pelo *site* Google Books Ngrams, que também permite que o usuário faça buscas e produza gráficos de uso desses bigramas. Assim, estritamente falando, não foi verificado diretamente um *corpus* de textos, pois os textos das publicações indexadas pelo Google Books não são disponibilizados para os usuários. O Google Books disponibiliza apenas os bigramas. Assim, os dados foram os bigramas, juntamente com sua frequência de ocorrência nas publicações em inglês indexadas entre 1800 e 2008.

Ferramentas de análise

Foram empregadas as seguintes ferramentas nesta pesquisa: a. Interface da Brigham Young University para o Google Books N-Grams. Essa interface auxilia nas buscas no Google Books N-Grams. b. Visualizador Google Books N-Gram Viewer. Essa interface *on-line* permite a visualização em forma de gráfico das ocorrências do n-gramas na base de dados Google Books N-Gram. c. Etiquetador semântico USAS. Esse etiquetador *on-line* atribui etiquetas semânticas a cada item lexical submetido. d. Listas de valoração de *sentiment analysis* de Hamilton, Leskovec e Jurafsky (2016). Essas listas foram disponibilizadas por Hamilton, Leskovec e Jurafsky (2016) na *web*. Esta contém a avaliação da polaridade (positiva ou negativa) de milhares de itens lexicais, distribuídos nas décadas em que ocorreram nos *corpora* pesquisados por esses autores. e. *Script* desenvolvido exclusivamente para execução desta pesquisa. Esse *script*, escrito em modo Unix Shell, executou todo o trabalho de preparação e de processamento dos dados.

Termos de busca

Foram empregados 20 termos de identificação do ser humano na pesquisa, quais sejam: *adolescent(s), adult(s), boy(s), child/children, girl(s), kid(s), man/men, teen(s), teenager(s), woman/women*. Esses termos foram selecionados porque indicam de algum modo as fases da vida em inglês.

1. These methodological innovations have enabled researchers to ask fundamentally different kinds of research questions.

2. A collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. (SINCLAIR, 2005, p. 4)

3. Social representations concern the contents of everyday thinking and the stock of ideas that gives coherence to our religious beliefs, political ideas and the connections we create as spontaneously as we breathe. They make it possible for us to classify persons and objects, to compare and explain behaviors and to objectify them as parts of our social setting. While representations are often to be located in the minds of men and women, they can just as often be found 'in the world', and as such examined separately..

Foram consultados dicionários e *thesauruses* para auxiliar na seleção dos termos. Foram feitos estudos pilotos para averiguar até que ponto os termos eram informativos. Vários termos foram descartados porque não se mostraram informativos, como *"elderly"*. Mesmo com a seleção final dos termos, ainda há termos que não são específicos de fase de vida, como *"man"*, que identifica o próprio ser humano. Essa ambiguidade de alguns termos é uma das limitações do trabalho. No entanto, ao olhar os colocados desses termos por meio dos bigramas, o termo pode ser melhor especificado e, por conseguinte a ambiguidade pode ser reduzida. Por colocados, ou *collocations*, é possível verificar que tratam de palavras que, ao serem utilizadas em conjunto, naturalmente, produzem um significado único. Não se trata de unir qualquer nódulo de palavra que faça um aparente sentido gramatical e, sim, palavras que se atraem naturalmente. Os termos foram divididos em grupos etários e de gênero, conforme mostra o quadro a seguir, para a análise das representações:

Quadro 1. Termos de busca empregados no estudo

Termos relacionados à infância
Termos não-marcados por gênero: *child;children;kid;kids.*
Termos femininos: *girl;girls.*
Termos masculinos: *boy;boys.*
Termos relacionados à adolescência:
Termos não-marcados por gênero: *adolescent;adolescents;teen;teens; teenager; teenagers.*
Termos relacionados à idade adulta
Termos não-marcados por gênero: *adult;adults.*
Termos femininos: *woman; women.*
Termos masculinos: *man; men.*

Fonte: Elaboração própria

Interface do Google Books N-gram Database da Brigham Young University

Conforme mencionado, a Brigham Young University oferece uma interface de busca para a base de dados Google Books N-Grams; e é oferecida a opção de escolher uma das três bases de dados relativas ao Google Books N-Grams: American (*155 billion words*), British (*34 billion words*) e Spanish (*45 billion words*). Cada uma dessas bases é etiquetada gramaticalmente, o que permitiu a busca pelos colocados adjetivos. Escolhemos a opção "American English" por ser a maior base de dados. Não há a opção de buscar toda a base de dados do inglês.

Análise das representações

Para identificar as representações latentes nos bigramas, foram empregadas metodologias quantitativas e qualitativas. A metodologia quantitativa empregou a etiquetagem semântica descrita acima bem como o processamento do *script* que identificou as cinco classes semânticas mais ocorrentes dos colocados. Com base nas categorias semânticas mais recorrentes, foi possível visualizar as categorias semânticas presentes nos dados estudados. A partir dessa primeira impressão, os colocados foram classificados de modo qualitativo em categorias que, a nosso ver, poderiam revelar as representações latentes nos dados. A análise das representações teve como fundamentação Moscovici (2000), Baker (2013) e Berber Sardinha (2014). A seguir, relata-se os principais resultados da análise de dados encontrados nesta pesquisa de Doutorado.

Análise dos dados

Serão descritos aqui os principais resultados desta pesquisa, a fim de responder as perguntas do estudo. Por razões de limitações de espaço via artigo, os resultados serão resumidos aos termos relacionados à infância e mais adiante dados referentes às outras faixas etárias também estudadas.

Termos relacionados à Infância

A categoria semântica mais frequente é a baseada na análise de sentimento executada em *script*, relativa a GENERAL & ABSTRACT TERMS (Termos Gerais e Abstratos), que se trata de verificar aqui a análise de emoções *per se.* Trata, portanto, do método que classifica emoções específicas, como ódio, amor ou felicidade, presente nesta mensagem, Portanto, a mais frequente, com 18 colocados como 'average', 'defective', 'dependent', 'disadvantaged', 'exceptional', 'good', 'hyperactive', 'mere', 'minor', 'natural', 'normal', 'other', 'particular', 'perfect', 'real', 'specific', 'true', 'typical'. A segunda categoria é a T, relativa a TIME (período de tempo), com 10 colocados como 'eldest', 'modern', 'new', 'newborn', 'old', 'older', 'oldest', 'young', 'younger', 'youngest'. A terceira categoria é a N, relativa a

NUMBERS & MEASUREMENT (Números & Métrica), com 10 colocados como 'additional', 'eighth', 'fourth', 'ninth', 'seventh', 'single', 'small', 'tenth', 'tiny', 'whole'. A quarta categoria é a E, relativa a EMOTIONAL ACTIONS, STATES & PROCESSES (Ações, Estados e Processos Emocionais), com 10 colocados como 'aggressive', 'battered', 'beloved', 'dear', 'dearest', 'favorite', 'happy', 'precious', 'shy', 'unhappy'.

A partir dessas categorias, podemos sugerir as representações de *child* como sendo as principais: (1) avaliatividade-normatividade (*beloved, dear, dearest, defective, dependent, exceptional, favorite, good, natural, new, perfect, precious, real, specific, whole)*; (2) quantificação *(additional, eighth, eldest, fourth, ninth, seventh, single, tenth*); (3) gradação etária (*minor, newborn, old, older, oldest, young, younger, youngest*); (4) comportamento (*aggressive, happy, shy, unhappy*); (5) normatividade (*average, normal, typical*). Em suma, '*child*' é uma figura que é fundamentalmente avaliada, contada e mensurada. Em relação à variação temporal, os números indicam que tanto o aumento quanto a diminuição estão relacionados a uma representação de cunho avaliativo. No entanto, o aumento está relacionado a dois aspectos da gradação do conceito de criança: *small child e older/ youngest child*, que praticamente não existiam no início do século XIX. Isso sugere que o conceito de 'criança' mudou de algo monolítico para algo gradativo em termos de tamanho e/ou de idade. Também houve um acréscimo do conceito de criança relacionado a aborto (*unborn child*) e a questões físicas (*handicapped child*). Por outro lado, houve uma diminuição da representação da criança como uma criatura inocente, como 'innocent child', 'beloved child', 'sweet child'. Mesmo 'poor child' remete a um juízo de valor nem sempre relacionado à condição financeira.

Desta maneira, os resultados da análise de representações mostram três padrões drasticamente diferentes dos três grupos. Na infância, são preponderantes as representações físicas (*barefoot, beardless, beautiful, bigger, gallant, handsome, large, larger, small, smaller*). Há ainda representações de comportamento (*aggressive, happy, idle, merry, nice, mischievous, rough, rude, wanton*), da própria idade (*oldest, preschool, senior, teenage, youngest*) e de superioridade/virtude (*popular, promising, smart, stable*).

Na adolescência, as principais representações são comportamentais (*angry, bored, noisy, restless, rowdy, sullen, violent*), sociais (*disadvantaged, drunk, drunken, runaway, unemployed, vulnerable*), clínicas (*deaf, depressed, disabled, handicapped, ill, suicidal*), de identificação de gênero (*bisexual, female, lesbian, male*) e de inferioridade (*impressionable, inexperienced, troubled*).

Na idade adulta, a representação dominante é de superioridade/sucesso (*ambitious, civilized, distinguished, eminent, famous, great, greatest, honest, illustrious, important, influential, mighty, powerful, principal, prominent, reasonable, remarkable, sensible, successful, thoughtful, true, wise, wisest, worthy*), seguida de representações de ocupação (*literary, medical, military, professional, public, scientific)* e de inferioridade (*desperate, lesser, primitive, strange, wicked*).

Assim, com base nessas categorias, parece haver um padrão de representação das três fases que segue uma trajetória que vai da aparência física, do comportamento e da classificação etária, na infância, para questões comportamentais, clínicas, de identificação de gênero e de inferioridade, na adolescência, para uma representação de sucesso e superioridade, na vida adulta. Há, portanto, fortes indícios de uma valorização da infância e especialmente da vida adulta, em detrimento da adolescência.

Considerações finais

A presente pesquisa teve como objetivo identificar as representações associadas a termos que designam o ser humano em inglês, a partir da utilização da base de dados Google Books NGrams, cobrindo um período de tempo que vai do início do século XIX ao início do século XXI.

Um total de vinte termos foram investigados, divididos entre termos relacionados à infância, femininos (*girl, girls*), masculinos (*boy, boys*) e não-marcados por gênero (*child, children, kid, kids*); termos relacionados à adolescência (todos não marcados por gênero, *adolescent, adolescents, teen, teens, teenager, teenagers*) e à idade adulta, femininos (woman, women), masculinos (*man, men*) e não marcados por gênero (*adult, adults*).

A pesquisa ofereceu, pela primeira vez, um panorama geral de como a vida humana é representada historicamente na língua inglesa, no tocante às suas fases etárias e diferenciação entre gêneros. Além disso, a pesquisa aqui relatada mostrou a impossibilidade de generalizações amplas: cada termo possui um leque próprio de representações, que o distingue dos outros termos, mesmo os mais próximos conceitualmente ou morfologicamente.

A língua em uso resiste a generalizações amplas: há muitas nuances entre os termos. A generalização que podemos fazer, com base nos resultados, é que a passagem da vida é marcada por uma constante classificação do ser humano em termos de um conjunto mais ou menos finito de representações, com viés notadamente avaliativo e normativo. E que essa passagem é marcada historicamente, com mudanças temporais marcantes nos últimos 200 anos.

Agradecimentos

Quero agradecer muito em primeiro lugar à equipe do GEL, a qual participei de submissão de trabalhos em 2017, mas neste momento em 2020, tornei-me membro com a anuidade. Respeito muito este grupo e também não posso deixar de agradecer ao meu orientador do Mestrado e Doutorado, Professor Doutor Tony Berber Sardinha, quem pacientemente por todos estes anos (2011-2019) me orientou e ao grupo de pesquisas de Linguística de *Corpus* da PUCSP, o GELC.

REFERÊNCIAS

ALVAREZ, M. L. O., SILVA, K. A. *Linguística aplicada:* múltiplos olhares. Brasília: UnB, Campinas: Pontes Editores, 2007.

BAKER, P. *Using Corpora to Analyze Gender.* London: Bloomsbury, 2014.

BAKER, P.; ELLECE, S. *Key Terms in Discourse Analysis*. London: Continuum, 2014.

BAKER, P.; GABRIELATOS, C.; MCENERY. T. *Discourse Analysis and Media Attitudes:* The Representation of Islam in the British Press. Cambridge: Cambridge University Press, 2013.

BERBER SARDINHA, T. *On Being American and Brazilian in Google Books*: A multi-dimensional perspective. Looking at cultural shifts in English over time: A Multi-Dimensional perspective. American Association for Corpus Linguistics Conference, Flagstaff, AZ, 2014.

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

BERBER SARDINHA, T. Linguística de corpus: histórico e problemática. *Delta*, v. 16, n. 2, p. 62, 2000.

BIBER, D. *University language:* A corpus-based study of spoken and written registers. Amsterdam: John Benjamin's, 2010.

BIBER, D. *Variation Across Speech and Writing.* Cambridge: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S. *Register, genre, and style.* Cambridge: Cambridge University Press, 2009.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics:* investigating language structure and use. Cambridge: Cambridge University Press, 1998.

DUBOIS, J; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J. B.; MEVEL, J. P. *Dicionário de lingüística*. São Paulo: Cultrix, 1993.

GALISSON, R.; COSTE, D. *Dicionário de didáctica das línguas*. Coimbra: Livraria Almedina, 1983.

HAMILTON, J. L.; LESKOVEC, J.; JURAFSKY, D. *Word Embeddings Reveal*, St. Press, 2016.

KENNEDY G. *An introduction to corpus linguistics*. London: Longman; 1998.

LODGE, D. *Changing Places:* A Tale of Two Campused. England: Penguin Books, 1975.

MCENERY, T.; WILSON A. *Corpus linguistics.* Edinburgh: Edinburgh University Press; 1996.

MOSCOVICI, S. Notes towards a description of social representations. *European Journal of Social Psychology*, Paris, v. 18, p. 211-250, 1988.

MOSCOVICI, S. *Representações sociais:* investigações em psicologia social. Rio de Janeiro: Vozes, 2003. Tradução Pedrinho A. Guareschi, a partir do original em língua inglesa: DUVEEN, G. (ed.). *Social representations*: explorations in social psychology. Nova York: Polity Press/Blackwell Publishers, 2000.

SINCLAIR, J. McH. *Corpus, Concordance, Collocation*. Oxford, New York: Oxford University Press, 1991.

STUBBS, M. *"British traditions in text analysis: from Firth to Sinclair"*, Text and Technology: In Honour of John Sinclair, ed. M. Baker, G. Francis & E. Tognini-Bonelli. Amsterdam: John Benjamins, 1993. p. 1-33.

STUBBS, M. Text and Corpus Analysis Computer-assisted Studies. *In:* SINCLAIR, J. McH. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1993.

STUBBS, M. *Text and Corpus Analysis:* Computer-Assisted Studies of Language And Culture, (University of Trier) Blackwell Publishers (Language in Society series, edited by Peter Trudgill, v. 23), 1996,

TRASK, R. L. *Dicionário de Linguagem e Lingüística*. São Paulo: Contexto, 2004.