http://dx.doi.org/10.31513/linguistica.2020.v16n1a31709
Recebido em: 17/01/2020 – Aceito em: 18/04/2020

# Sobre a constituição de corpora para línguas com poucos recursos (*less-resourced languages*)[1]

## *On the constitution of corpora of less-resourced languages*

*Lílian Teixeira de Sousa*[2]

## RESUMO

O uso de *corpora* em estudos linguísticos é bastante antigo, já a área da Linguística de *Corpus* é relativamente nova, tendo sua origem vinculada à ampliação do acesso a computadores e, consequentemente, ao Processamento de Linguagem Natural (PLN). À medida que a área foi ganhando influência na pesquisa linguística, o conceito de *corpus* foi se tornando mais específico e elementos como amplitude e referência, além de legibilidade por máquina e tamanho finito, passaram a se tornar fundamentais para a composição de amostras na área. Ao mesmo tempo, no entanto, foram surgindo *corpora* menores e bem menos amplos constituídos com objetivos bastante distintos, como, por exemplo, para a realização de documentação de línguas ameaçadas. Partindo disso, o presente artigo tem por objetivo discutir as diferenças entre *corpora* "prototípicos" criados segundo os pressupostos da Linguística de *Corpus*, e os *corpora* de línguas com pouca presença digital (*less-resourced languages*). Mostro que os *corpora* de línguas com pouco recursos tendem a ser mais especializados e, dificilmente, cumprem todos os critérios exigidos de um corpus amplo e representativo de uma língua. Apesar dos limites impostos por questões específicas de cada língua, concluo que a constituição de *corpora* para línguas com poucos recursos, ainda que não cumpram todos os critérios propostos pela Linguística de *Corpus*, devem ser realizados, e os resultados devem ser aproveitados de diversas formas, seja gerando novas tecnologias, servindo de suporte empírico para teorias linguísticas ou promovendo a língua na comunidade.

**Palavras-chave:** línguas com poucos recursos, PLN, corpus

## ABSTRACT

While the use of corpora in linguistic studies is quite old, Corpus Linguistics is a relatively new area of study, emerging with the expansion of access to computers and, consequently, to Natural Language Processing (NLP). As the subject gained influence within linguistic research, the concept of corpus became more specific. Breadth of sampling and standard references, as well as machine readability and finiteness became essential elements to compose samples. At the same time, however, smaller and much narrower corpora emerged, having distinct purposes, such as those documenting endangered languages. From this understanding, this paper aims to discuss differences between "prototypical" corpora built from the assumptions of Corpus Linguistics and those of less-resourced languages, with a small digital footprint. I demonstrate that the corpora of less-resourced languages tend to be more specialized and hardly ever fulfill the criteria required of a broad and representative corpus. In spite of limitations entailed by issues specific to each language, I conclude that the constitution of corpora for less-resourced languages must be undertaken, even if they do not fulfill all desirable criteria of Corpus Linguistics. The results must be exploited in diverse ways, whether though the creation of new technologies, as empirical support for linguistic theories or in promoting the language in the community.

**Keywords:** less-resourced languages, NLP, corpus

1 Agradeço aos pareceristas desta revista pelos comentários e sugestões que contribuíram para a qualidade do texto.

2 Universidade Federal da Bahia (UFBA), Setor de Linguística. E-mail: lilian.sousa@ufba.br

## 1 Introdução

Se considerarmos o conceito de *corpus* como "um conjunto de textos escritos ou falados numa língua, disponível para análise" (TRASK, 2004, p 68-69), pode-se dizer que o uso de *corpora* em estudos linguísticos é bastante antigo, o Corpus Helenístico, por exemplo, data da Antiguidade (BERBER SARDINHA, 2004). Já no século XX, muitos pesquisadores se dedicaram à descrição linguística através de dados coletados para esse propósito. Nos dias de hoje, no entanto, com o uso do computador, a linguística empírica ganhou novos contornos, uma vez que mais pesquisadores passaram a ter acesso ao Processamento de Linguagem Natural[3] (PLN) e, consequentemente, ampliaram-se as possibilidades de criação e manutenção de *corpora* maiores e mais amplos. Com isso, a Linguística de *Corpus* ganhou influência na pesquisa linguística e passou a ter definições próprias. Berber Sardinha (2004, p 3), por exemplo, a define como uma "abordagem que se ocupa da coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística". Já McEnery e Wilson (2001) tratam a Linguística de *Corpus* não como uma área, mas como uma metodologia que tem por finalidade estudar a língua baseada em exemplos da vida real. Mesmo o conceito de *corpus* se tornou mais específico para dar conta das inovações promovidas pela tecnologia. Segundo Sinclair (2005), um *corpus* é uma coleção de textos de uma língua em formato eletrônico, selecionado por critérios externos para representar, tanto quanto possível, uma língua ou variedade como fonte de dados para a pesquisa linguística.

McEnery e Wilson (2001) tornam o uso do termo *corpus* ainda mais específico ao apresentar quatro características fundamentais, a saber, (i) *amostragem e representatividade* (*sampling and representativeness*), um *corpus* deve ter uma amostragem suficiente da língua ou variedade que se quer analisar para obter o máximo de representatividade; (ii) *tamanho finito* (*finite size*), um *corpus* é finito independentemente de sua extensão (e.g. em número de palavras); (iii) *formato eletrônico* (*machine-readable form*), deve estar em formato que possa ser processado pelo computador; e (iv) *referência padrão* (*standard reference*), um *corpus* deve ser uma referência padrão[4] da variedade de língua que representa. Segundo Aluísio e Almeida (2006), essa última característica é uma diferença marcante entre a concepção de *corpus* para a Linguística e para a Linguística de *Corpus*, já que a construção de um *corpus* nessa perspectiva não serve apenas para uma única pesquisa, mas pode ser

3 Processamento de linguagem natural é uma subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais.

4 Padrão aqui refere-se à regularidade expressa na recorrência sistemática de unidades coocorrentes de várias ordens (lexical, gramatical, sintática etc.)

útil para outros pesquisadores por se tratar de uma referência padrão de uma língua ou variedade, permitindo assim pesquisas sobre qualquer área da linguagem, de fonética a discurso. Assim, uma das principais características da Linguística de *Corpus* é tornar possível generalizações de um *corpus* para uma língua como um todo ou pelo menos para uma variedade ou registro particular.

Paralelo ao desenvolvimento da chamada Linguística de *Corpus*, que, como apontado acima, apresenta definições próprias de *corpora*, começaram a surgir programas de financiamento de projetos de documentação de línguas pouco estudadas ou ameaçadas no mundo todo que resultaram na construção de *corpora* menores e menos amplos (SCANNEL, 2007; OSTLER, 2008; COX, 2011). Isso porque muitos linguistas passaram a reconhecer a necessidade de se realizar a documentação de línguas minoritárias ou ameaçadas. Especialmente depois que Kraus (1992) afirmou que o próximo século veria a morte ou condenação de 90% das línguas humanas. Esses *corpora*, no entanto, não satisfazem todas as condições apontadas no parágrafo acima, já que um bom número de dados presentes nesses bancos foi produzido unicamente para esse fim, não se tratando de dados de contextos comunicativos reais, além do que muitas vezes não conseguem ser amplos o suficiente para serem representativos. Uma outra questão importante é de natureza cultural e ética, já que em sociedades de cultura oral pode haver resistência ao registro escrito e/ou audiovisual[5], especialmente em rituais sagrados, o que restringe os contextos de documentação.

A questão que surge, então, é sobre a funcionalidade de *corpora* eletrônicos de línguas pouco estudadas, já que, na maioria das vezes, esses *corpora* não apresentam todas as características apontadas como necessárias pelos teóricos da área. Assim, o objetivo desse artigo é discutir as diferenças entre *corpora* "prototípicos" e *corpora* eletrônicos de línguas com pouca presença digital, conhecidas na área de Linguística Computacional como *less-resourced languages* (JOKINEN, 2018; SZYMANSKI, 2011), buscando apontar sua importância e potencial.

O artigo está dividido da seguinte forma: na seção 2, são apresentadas e discutidas algumas definições importantes. Na seção 3, descrevo alguns *corpora* de línguas com pouco recursos e sua funcionalidade. Na seção 4, discuto a importância da criação desse tipo de *corpora* e, por fim, apresento, na seção 5, as Considerações Finais.

## 2 Algumas definições

Na seção anterior, argumentei que juntamente com o crescimento e fortalecimento da Linguística de *Corpus* e consequente criação de *corpora* amplos e representativos com uma variedade de dados

5 Agradeço a um dos pareceristas que me lembrou da importância de fazer referência às dificuldades de ordem cultural.

de língua em contextos comunicativos naturais, cresceu nas últimas décadas o número de *corpora* construídos via documentação linguística. A Documentação linguística é um campo de atuação da linguística que se ocupa com a criação de registros de línguas em uso, através da criação de acervos digitais que permitem seu acesso e o uso de dados organizados nessas plataformas. Dados coletados via projetos de documentação, no entanto, são bem diferentes do tipo de dado que está presente em *corpora* construídos a partir dos critérios da Linguística de *Corpus*, a começar pelo fato de que, em projetos de documentação, os dados, embora englobem vários tipos de gênero, são coletados para esse fim, já que muitas das línguas objeto de documentação são pouco utilizadas em meio digital, algumas ágrafas, enquanto que, em um *corpus* "prototípico", os dados que compõem a amostra foram produzidos com objetivos comunicativos variados em contextos reais de uso da língua. Assim, é preciso distinguir o que é um *corpus* produzido conforme os pressupostos da Linguística de *Corpus*.

Como descreve Sanchez, para a Linguística de *Corpus*, o conceito de *corpus* diz respeito a

> 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 linguístico 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. (SANCHEZ, 1995, p.8-9)

Dessa forma, vemos que para essa área o conceito de *corpus* vai muito além da definição geral de "conjunto de documentos que servem de base para a descrição ou estudo de um fenômeno" (Dicionário on-line Priberam, 2020). Segundo Aluísio e Almeida (2006), há um conjunto de requisitos que impactarão na validade e confiabilidade de uma pesquisa segundo os pressupostos da Linguística de *Corpus*; assim, para o projeto de um *corpus* computadorizado, é preciso observar seis características: autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho. Assim, *corpus* ideal para a Linguística de *Corpus*, deve conter:

1. textos autênticos, o que significa que os textos devem ter sido escritos em linguagem natural, não podendo ser textos produzidos com o propósito de serem alvo de pesquisa linguística e devem ser produzidos por falantes nativos;
2. representatividade, deve ser representativo de uma língua ou variedade de língua que se deseja pesquisar. Nesse sentido, ao selecionar os textos, deve-se perguntar: Quais tipos de textos? Quais gêneros textuais? E se de fato representam os usos linguísticos de uma comunidade;
3. balanceamento, o *corpus* deve ter um equilíbrio de gêneros discursivos (informativo, científico, religioso etc.) ou de tipos de textos (artigo, entrevista, dissertação, carta etc.), ou de títulos, ou de autores ou de todos esses itens em conjunto;

4. amostragem proporcional, o tamanho das subamostras (falantes, registros, variedades, etc.) deve ser proporcional à proporção de falantes, registros, variedade, etc.;
5. diversidade, se o *corpus* pretende representar/ ser representante de uma língua, ele deve apresentar diversidade dialetal, de tópicos, de gênero para permitir diferentes tipos de estudo;
6. tamanho adequado ao tipo de pesquisa que se vai realizar e à metodologia a ser adotada na pesquisa.

Em relação às etapas metodológicas, é preciso (i) selecionar os textos a serem compilados, que podem ser provenientes da web ou textos impressos[6] (e posteriormente digitalizados), converter em formato txt e, na compilação, seguir as normas legais de anonimato e para a obtenção de direitos autorais; (ii) estabelecer os níveis de representação das informações presentes num *corpus*: anotação estrutural e anotação linguística. A anotação estrutural corresponde à marcação de dados externos (metadados textuais) e internos (segmentação – capítulos, parágrafos, etc.) dos textos. Já a anotação linguística trata dos níveis que se deseja representar (morfossintático, sintático, semântico, discursivo, etc.), que podem ser inseridos manualmente (por linguistas), automaticamente (por ferramentas de PLN) ou semi-automaticamente (correção manual de saída de outras ferramentas).

Considerando todas as características apresentadas de 1 a 6, muitos *corpora* importantes disponíveis não poderiam ser definidos como prototípicos. Como a Linguística de *Corpus* lida com textos produzidos com objetivos comunicativos reais, é preciso que haja textos disponíveis em diferentes gêneros e tipologias e com possibilidade de liberação de direitos autorais. Assim, se o objetivo é construir um *corpus* de uma língua com grande presença digital, é possível selecionar textos de diferentes gêneros e tipologias, já em formato digital e de acesso livre (blogs, postagens em mídias sociais, artigos etc.). Se, ao contrário, a língua objeto tem pouca presença digital, ou se está restrita a apenas alguns contextos comunicativos, os critérios de diversidade, balanceamento e mesmo de amostragem proporcional não são cumpridos. Vale lembrar que, em sociedades multilíngues, há falantes que distinguem os contextos comunicativos em que usam uma ou outra língua, ficando algumas línguas restritas ao contexto doméstico e outras às trocas comerciais. Também nos *corpora* históricos, só é possível encontrar as seis características em *corpora* de línguas com uma longa tradição escrita, o que acaba restringindo os *corpora* a algumas línguas europeias, asiáticas e ao árabe, já que muitas línguas são ágrafas, ou criaram muito recentemente um sistema de escrita, ou surgiram há pouco tempo. Em relação às línguas indígenas brasileiras, por exemplo, há pouquíssimos registros históricos escritos e em apenas alguns poucos gêneros – gramáticas (BATISTA, 2005) catecismos, dicionários

6 As tecnologias de processamento de fala ainda têm dificuldade de lidar com diferentes frequências de voz, dialetos e outros elementos difíceis de parametrizar. Por isso, mesmo os textos orais são antes transcritos.

(Enciclopédia das línguas no Brasil, 2020).

Já o critério de tamanho se relaciona ao critério de representatividade, embora isso seja bastante relativo, uma vez que que depende muito dos objetivos de pesquisa. Assim, é possível que um *corpus* seja representativo mesmo sendo de tamanho pequeno. Apenas a título de ilustração, apresento na Tabela 1 uma escala de tamanho proposta por Berber Sardinha (2002, p.119) baseada na análise de trabalhos completos publicados em cinco eventos dedicados à Linguística de *Corpus* (três edições do ICAME – *International Computer Archive of Mordern and Medieval English*, uma edição do PALC – *Practical Appications in Language and Computers* e uma edição do TALC – *Teaching and Language Corpora*):

**Tabela 1:** Classificação de *corpora* de acordo com o número de palavras

| **Classificação do *corpus*** | **Número de palavras** |
|---|---|
| Pequeno | Até 80 mil |
| Pequeno-médio | 80 a 250 mil |
| Médio | 250 mil a 1 milhão |
| Médio-grande | 1 milhão a 10 milhões |
| Grande | Acima de 10 milhões |

Fonte: Berber Sardinha (2002, p.119)

A presença ou não de todas as características de 1 a 6 distingue basicamente dois tipos de *corpora*: (i) os gerais, que pretendem representar a língua de forma ampla e servir de base para pesquisas variadas, caracterizando-se pela variedade de gêneros discursivos, registros, assuntos e autores; e (ii) os especializados, coletados para objetivos específicos de pesquisa e que consistem em coleções de textos de gêneros ou discursos específicos. Como um exemplo de *corpus* geral temos o *British National Corpus* (BNC), criado nos anos 80, que contém 100 milhões de palavras de textos de uma ampla variedade de gêneros já anotados e etiquetados gramaticalmente. Há, claro, *corpora* gerais de várias outras línguas: francês (SIEPMANN; BÜRGEL, 2014), espanhol (SANCHEZ *et al.,* 1995), árabe (ARTS *et al.,* 2014), chinês (ZHOU; YOU, 1997), estoniano (HENNOSTE *et al.,* 1998), etc. No Brasil, o desenvolvimento da área da Linguística de *Corpus* é mais recente e muitos dos *corpora* compilados são especializados, como o do Projeto DIRECT (PUC-SP), composto de textos de comunicação no contexto de negócios, e do Projeto NURC, composto de textos de fala culta de diferentes regiões urbanas do país, colhidas em situações pré-estabelecidas (OLIVEIRA, 2016). Dentre os *corpora* gerais, podemos citar o CORPOBRAS - PUC-Rio (OLIVEIRA; DIAS, 2009), que busca ser representativo do português do Brasil, apresentando 27 gêneros discursivos e mais de 1.000.000 de palavras; e o *Corpus* Anotado do Português Histórico - Tycho Brahe (Unicamp), com mais de 3.300.000 palavras.

Em relação a ferramentas computacionais, é importante dizer que todos os *corpora* gerais listados no parágrafo acima têm vários tipos de anotação que permitem a aplicação ou desenvolvimento de diversas ferramentas computacionais de busca. Tomando o COCA – *Corpus of Contemporary American English* como exemplo, por ser um *corpus* bastante amplo, com mais de 1 bilhão de palavras, e balanceado, com oito diferentes gêneros, vemos que o sistema sofisticado de anotação permite vários tipos de busca – por palavras, sintagmas, subsequências, lexemas, morfemas, sinônimos, listas de palavras. Com isso, é possível realizar vários tipos de pesquisas, através da frequência de palavras, sintagmas e construções gramaticais, distinguindo gênero textual e período histórico; é possível, por exemplo, comparar a colocação de palavras mais frequentes em um gênero e outro. Também é possível mapear mudanças em construções sintáticas comparando a frequência e uso dessas construções considerando os textos mais recentes e os mais antigos.

Para línguas minoritárias ou ameaçadas é ainda mais difícil encontrar um *corpus* que cumpra mais do que três dos critérios apontados. É importante lembrar que os *corpora* de línguas ameaçadas são, muitas vezes, oriundos de projetos de documentação, que são sempre menores do que os de línguas de maior *status*, e que não podem, geralmente, ser considerados representativos de um tipo particular de falante, registro e nem mesmo variedade, além de, algumas vezes, serem de acesso restrito aos que participaram de sua criação. É o caso dos projetos de documentação realizados a partir do Programa DOBES que compõem o *The Language Archive* (TLA) localizado no Instituto Max Planck de Psicolinguística em Nijmegen[7] (https://dobes.mpi.nl/projects/?lang=pt). Essa grande base de dados apresenta quatro níveis de acesso: no nível 1, há arquivos que podem ser abertos por qualquer pessoa; no nível 2, há arquivos que podem ser acessados por usuários registrados que assinaram o Código de Conduta. No nível 3, há arquivos disponíveis apenas para usuários aprovados por um contribuinte e, no nível 4, os arquivos são fechados (acessados apenas pelo(s) pesquisador(es) que realizaram a documentação).

No Brasil, há várias iniciativas de documentação linguística e cultural realizadas com as comunidades indígenas, que resultaram em grandes acervos digitais armazenados no Museu do Índio, no Rio de Janeiro, e no Museu Paraense Emílio Goeldi – MPEG (https://www.museu-goeldi.br), em Belém do Pará. O MPEG é pioneiro no trabalho de documentação linguística no Brasil, especialmente na documentação das línguas faladas na Amazônia. Segundo Galúcio (2004), o acervo do MPEG possui material de mais de 50 línguas, algumas dessas extremamente ameaçadas. Já o Museu do Índio

7 O TLA é especializado no arquivamento e na preservação de gravações linguísticas e de outros tipos de dados, bem como no desenvolvimento de ferramentas linguísticas e de seu arquivamento. Já o programa DOBES (em alemão *"Dokumentation bedrohter Sprachen"*, ou seja "Documentação de línguas ameaçadas") foi iniciado pela Fundação Volkswagen com o propósito de documentar línguas potencialmente sob perigo de extinção dentro de poucos anos.

possui o maior acervo digital de línguas indígenas do Brasil graças à iniciativa da criação do Programa de Documentação de Línguas e Culturas Indígenas – PROGDOC (http://progdoc.museudoindio.gov.br). No entanto, o único *corpus* de língua indígena brasileira de livre acesso e anotado é do kadiwéu vinculado ao Tycho Brahe, na Unicamp (GALVES; SANDALO; SENA; VERONESI, 2017). O projeto kadiwéu é pioneiro ao disponibilizar um *corpus* anotado de língua indígena americana que permite pesquisas gramaticais que incluem informação morfológica e sintática.

De maneira geral, há muitas línguas que não dispõem de qualquer tipo de *corpus* eletrônico. Para se ter uma ideia da pouca amplitude dos *corpora* que temos hoje em relação à quantidade de línguas humanas, basta observar que, embora a maior parte seja de línguas indo-europeias, das 6000 línguas vivas estimadas, 3000, ou seja 50%, são faladas na Ásia, por volta de 1900 na África (31%), 900 nas Américas (15%) e somente 275 são faladas na Europa e no Oriente Médio, o que corresponde a apenas 4%. Ainda mais revelador é que a maior parte dessas línguas está concentrada em 22 países: Papua Nova Guiné (850), Indonésia (670), Nigéria (410), Índia (380), Camarões (270), Austrália (250), México (240), Congo (210) e Brasil (160).

Toda essa discussão em torno da composição da documentação de línguas ameaçadas leva a um outro ponto importante a ser abordado nessa seção; me refiro ao uso do termo *less-resourced languages*. O termo presente no título, como mencionado, tem sua origem na Linguística Computacional e é usado para definir línguas em que falta uma presença digital significativa (SZYMANSKI, 2011).

Eu também poderia ter escolhido os termos 'línguas pouco estudadas', 'ameaçadas' ou 'minoritárias', mas o fato é que, dentro do conceito de *less-resourced languages*, estão incluídas, na maioria das vezes, línguas minoritárias ou ameaçadas, mas não apenas. Há casos de línguas que não podem ser classificadas como ameaçadas, mas que, ainda assim, apresentam pouca presença digital, o que pode ser explicado pela ausência de uma tradição escrita, ou mesmo de uma grafia, e pela falta de acesso a recursos tecnológicos por parte dos falantes. Já o termo 'línguas pouco estudadas' é muitas vezes equalizado a *less-resourced languages*, uma vez que na prática a ausência de recursos significa também a ausência de estudos científicos, pelo menos na área de PLN. A escolha entre esses dois termos, deu-se por dois fatores, pela existência de uma referência na área da Linguística Computacional e também porque, neste artigo, advogo justamente pela criação de *corpora*, ainda que não possam ser representativos da totalidade de usos linguísticos de uma língua.

Ainda tratando das definições, Maxwell e Hughes (2006, p.30) preferem o termo *lower density languages*. Segundo os autores, há apenas um pequeno número de línguas do mundo, como o inglês,

o chinês, o árabe e línguas da Europa Ocidental, cujos recursos são abundantes e que poderiam, por isso, serem definidas como línguas de alta densidade. Para algumas outras poucas, os recursos, se não abundantes, pelo menos existem e crescem cada vez mais; essas seriam as línguas denominadas como de densidade média, que é o caso das demais línguas europeias. Ainda segundo os autores, esses dois conjuntos somados não chegam a representar 30% das línguas do mundo. Isso significa que, para a grande maioria das línguas, os recursos são escassos.

Se, como afirmamos acima, a ausência de recursos implica a ausência de estudos científicos, isso significa que a maioria dos estudos linguísticos empíricos está disponível apenas para 30% das línguas humanas. Se faltam recursos para o desenvolvimento de estudos empíricos, consequentemente, boa parte de nosso conhecimento científico linguístico fica restrito a um número muito pequeno de línguas. Tal fato leva a crer que, por um lado, a criação de *corpora* poderia ampliar consideravelmente as possibilidades de análise linguística; por outro, lembramos que a amplitude de dados proposta pelos teóricos da Linguística de *Corpus* tem a ambição de possibilitar generalizações do *corpus* para a língua, o que muitas vezes não é possível para línguas com poucos recursos. Assim, qual seria a finalidade da composição de *corpora* para essas línguas? Para responder a essa questão, apresento, nas seções seguintes, algumas discussões sobre a constituição de *corpora* de línguas com poucos recursos.

## 3 Sobre *corpora* de línguas com poucos recursos

Como mencionado, têm surgido nas últimas décadas vários *corpora* de línguas ameaçadas. O crescimento deste tipo de *corpus* está relacionado à reação por parte da comunidade acadêmica à eminência de extinção de boa parte das línguas humanas. Para se ter uma ideia da magnitude do problema, mesmo as estimativas mais otimistas projetam a perda de 50% das línguas do mundo no século XXI (KRAUS, 1992). Assim, surgiram, desde a década de 1990, vários programas de documentação de línguas ameaçadas, por exemplo, o *Dokumentation Bedrohter Sprachen* (DOBES), financiado pela Fundação *Volkswagem* na Alemanha (2000-2013), e o *Endangered Languages Documentation Programme* – ELDP (https://www.eldp.net), financiado pela *US National Science Foundation* (NSF) e pelo *National Endowement of the Humanities* (2005-).

Todo esse processo fez com que a Documentação Linguística, historicamente relacionada à criação de gramáticas, dicionários e coleção de textos, alcançasse o *status* de uma subárea da Linguística. Com isso, os métodos para gravação e análise de materiais linguísticos e culturais foram aprimorados, incluindo-se a criação de bancos de metadados e o uso de tecnologias computacionais.

Um dos pontos de divergência da metodologia de Documentação Linguística em relação à Linguística de *Corpus* é que, enquanto na documentação os dados são, muitas vezes, produzidos especialmente para serem incluídos no *corpus*, para os propósitos da Linguística de *Corpus* é essencial que os dados sejam reais, produzidos com objetivos comunicativos próprios da situação comunicativa real.

Assim, vemos que, embora a composição dos *corpora* coletados através de projetos de documentação seja diferente daquela dos *corpora* formados a partir da metodologia da Linguística de *Corpus*, é relevante dizer que eles existem e são usados na descrição e análise linguísticas. Já para um grande número de línguas não há qualquer tipo de *corpus*, isso porque somente 1% das línguas do mundo tem a vantagem de serem reconhecidas como línguas oficiais, apesar de seus falantes chegarem a compor 60% da população mundial. Além da população de falantes, o *status* dessas línguas também é desigual, o que também tem impacto sobre como *corpora* são criados e usados. McEnery e Ostler (2000, p. 405) apontam pelo menos oito classes de línguas que têm problemas para encontrar apoio governamental para realizar a construção de *corpus*:

(1) Línguas regionais oficiais sem status nacional (e.g. galês);
(2) Dialetos – variantes de uma ampla língua nacional (e.g. variedades do chinês);
(3) Vernáculos – línguas com um grande número de falantes sem reconhecimento oficial (e.g. Sylhet em Bangladesh) ;
(4) Línguas nacionais com uma população pequena de falantes (e.g. islandês);
(5) Línguas minoritárias – vernáculos com um número relativamente pequeno de falantes (e.g. a língua gaélica da Ilha de Man, no Reino Unido – manês);
(6) Línguas minoritárias não-indígenas, é o caso das línguas de imigrantes (e.g. Punjabi, no Reino Unido)
(7) Línguas ameaçadas (e.g. Kereque, com 100 falantes);
(8) Línguas de sinais (e.g. LIBRAS).

A esta lista, eu acrescentaria (9) Línguas extintas (e.g. latim, tupi), uma vez que há muitas línguas já extintas que apresentam, porém, textos escritos que poderiam compor um *corpus*. Nesse caso, também há particularidades, já que é praticamente impossível encontrar *corpora* históricos amplos e balanceados.

É importante retomar que a característica fundamental para a definição de línguas com poucos recursos é a ausência de recursos computacionais disponíveis, muito mais do que número de falantes ou quantidade de pesquisas realizadas. Assim, a dificuldade fundamental para o uso mais amplo de

*corpora* de línguas ameaçadas ou minoritárias está nos problemas com a anotação, que é ausente em alguns *corpora* ou deficiente em outros. Aqui é importante mencionar que a anotação é essencial para a pesquisa linguística envolvendo *corpora*, uma vez que *corpora* anotados podem ser usados para treinar algoritmos através de aprendizagem de máquina, o que permite a criação de recursos computacionais.

As possibilidades de anotação e geração de recursos é bastante ampla, há (i) texto paralelo alinhado com outra língua no nível sentencial e, às vezes, com um ou mais níveis de paralelismo; (ii) texto anotado por entidade nomeada em vários níveis de granularidade; (iii) textos analisados morfologicamente, também com um esquema de rótulos morfológicos apropriados para uma língua em particular; (iv) textos marcados por fronteiras de palavras, importante especialmente para línguas que não marcam a maioria das fronteiras de palavras; (v) texto com rótulo POS (*part of speech*) e com um esquema de rótulos POS apropriados para a língua particular (e.g. N (*noun* - substantivo), V (*verb* - verbo) , P (*preposition* - preposição etc.) ; (vi) banco de árvores (sintaticamente anotadas e parseadas); (vii) texto rotulado semanticamente e (viii) dicionários eletrônicos e outros recursos lexicais, como Wordnet[8].

Voltando às diferenças para a composição de bancos de dados entre a Linguística de *Corpus* e a Documentação Linguística, é relevante dizer que muitos *corpora* de documentação são de acesso restrito aos pesquisadores que atuaram em sua criação; o que torna impossível apresentar nessa seção uma análise completa da organização e funcionalidade de *corpora* de dessas línguas. Por esse motivo, minha argumentação parte de um tipo específico de *corpora* de línguas com poucos recursos, me refiro a *corpora* de tamanhos entre médio e grande de línguas conhecidas, mas com baixa presença digital. Assim, selecionei dois *corpora* sem qualquer tipo de anotação e dois anotados[9] para dar ao leitor uma ideia geral da diferença desses *corpora* em relação aos prototípicos.

Dentre *corpora* não-anotados, Vinogradov (2016) cita o Assamese e o Ndebele. O assamese, segundo o autor, é uma língua indo-iraniana falada por quase 13 milhões de pessoas na Índia. O *corpus* do assamese é composto de 3 milhões de *tokens* (palavras, números etc) de 1191 textos ao todo e está integrado ao Projeto EMILLE – Enabling Minority Language Engineering. Os textos são de um conjunto bastante variado de gêneros, mas, como não é anotado, a única forma de pesquisa no *corpus* é através da forma exata em que as palavras estão escritas. Já o ndebele é uma língua africana vinculada ao grupo banto da macro-família Níger-Congo e falada por aproximadamente um milhão

8 http://wordnet.princeton.edu

9 Consideramos como anotado mesmo aquele corpus em que há apenas um conjunto de dados anotados.

e meio de pessoas. O *corpus* dessa língua, composto por textos escritos e falados, contém 691.268 *tokens* e foi desenvolvido como parte do Projeto ALLEX – African Languages Lexicon. Esse último, no entanto, é voltado para a produção de dicionários e outras ferramentas para línguas africanas usadas no Zimbabué, o que o caracteriza como um *corpus* especializado.

Os dois *corpora* citados acima são de tamanho, respectivamente, médio-grande e médio, mas enquanto o *corpus* do assamese apresenta uma grande variedade de gêneros, o do ndebele distingue apenas entre textos escritos e falados, sem mencionar gênero. Nesse caso são dois *corpora* de bom tamanho e um deles bastante diverso, mas não contam com nenhum sistema de anotação.

Para seguir o mesmo padrão, selecionei também para a descrição dos *corpora* anotados uma língua iraniana, o osseto, e uma africana, o bambara. O osseto é falado por aproximadamente meio milhão de pessoas na parte central do Cáucaso (Federação Russa e Geórgia). O *corpus* do osseto é composto de textos escritos e contém mais de 11 milhões de *tokens*. Como é anotado, inclui informações gramaticais sobre os *tokens* e tradução. Segundo Vinogradov (2016), o principal mérito desse *corpus* está no mecanismo de busca e na interface amigável. Nesse *corpus*, é possível fazer buscas por lexema, forma da palavra, tradução ou por um conjunto de traços gramaticais. Também é possível selecionar mais de um *token* e checar a distância entre eles.

O bambara é uma língua Mandê que, como o ndebele, pertence à macro-família Níger-Congo. Esse é um dos poucos casos de língua com poucos recursos, mas não ameaçada, já que é falada por mais de 10 milhões de pessoas. O *corpus* do bambara é composto de 426.813 *tokens* de textos de diferentes gêneros e zonas dialetais. Como também é anotado e contém subcorpora desambiguados, permite ao usuário a pesquisa por lexema, forma das palavras, sintagmas, símbolos, classe de palavras, além de especificar o contexto e permitir a visualização dos resultados em diagramas de frequência.

A partir da descrição desses quatro *corpora*, é possível observar que as únicas duas características comuns entre eles são a legibilidade por máquina e a finitude, já que esses bancos de dados diferem quanto à quantidade e diversidade dos textos disponíveis para cada língua em particular. Em alguns casos, há uma drástica diferença na distribuição dos textos de acordo com o gênero, em outros, os dados não são autênticos, tendo sido produzidos via elicitação. Assim, vemos que não há uma uniformidade de tipos de dados e recursos entre os *corpora* dessas línguas, uma vez que a motivação para a sua criação não é sempre a mesma e eles tendem a ser mais especializados do que gerais.

Há ainda outro ponto que chama a atenção, são as diferenças em relação à possibilidade de uso de ferramentas computacionais. Os *corpora* não anotados, é claro, só permitem pesquisas

simples através da forma das palavras, mas mesmo entre os dois *corpora* anotados descritos acima, há diferenças quanto ao tipo de anotação realizado. Em alguns casos, o padrão de anotação permite uma grande variedade de buscas, o que o torna uma poderosa ferramenta para a descrição e análise linguísticas, mas ainda assim, as possibilidades de descrição e análise são bem distantes do que está disponível para os *corpora* de línguas com ampla presença digital. Relevante nesse caso é a quantidade de textos escritos produzidos. Se uma língua é ágrafa, mesmo sendo falada por um grande número de indivíduos, isso significa certamente que haverá pouca presença digital.

As dificuldades enfrentadas na documentação e criação de *corpora* de línguas ágrafas são tão reconhecidas que houve recentemente a criação de um projeto voltado para esse fim. O *Breaking the Unwritten Language Barrier* (BULB), como é chamado, aproxima linguistas e cientistas computacionais com o objetivo de desenvolver ferramentas que sirvam de auxílio para o trabalho de documentação de línguas ágrafas, como reconhecimento automático de fala e tradução por máquina (ADDA; ADDA-DECKER; AMBOUROUE; BESACIER; BLACHON *et al*, 2016).

Há ainda outras razões para a baixa frequência de textos eletrônicos (McENERY; OSTLER, 2000, p. 411): (1) baixo nível de letramento da comunidade; (2) um sistema de escrita ainda não computadorizado; e (3) um sistema de escrita computadorizado, mas sujeito à competição com padrões.

Em relação aos problemas de padronização, o Unicode[10], que se refere a um conjunto universal de caracteres para o sistema de escrita humana, é apontado como uma esperança para se resolver a questão, embora não seja capaz de sozinho resolver todos os problemas para a composição de *corpora*.

Há ainda outros problemas de ordem tecnológica enfrentados para a constituição de *corpora* eletrônicos para línguas com poucos recursos. As ferramentas e técnicas de processamento de texto existentes foram em sua maioria desenvolvidas a partir de aprendizagem de máquina, que precisa de uma grande quantidade de dados. Na falta desses, muitos programadores optam por sistemas de regras, o que, na verdade, pouco auxilia o trabalho do linguista que vai trabalhar com línguas pouco estudadas, já que para se criar ferramentas via sistema de regras é necessário já conhecer de antemão o funcionamento da língua.

Resumindo, quando se compara os processos de composição de *corpora* de línguas com poucos recursos com os considerados prototípicos, observa-se que os primeiros são mais limitados, alguns

10 Veja http://www.unicode.org/unicode/standard/standard.html para mais detalhes.

não são bem distribuídos em termos de gêneros e variedades, outros são bem distribuídos, mas não são anotados. Assim, vale refletir sobre que contribuições a constituição de bancos de dados de línguas com poucos recursos podem trazer, já que muitas vezes não são amplos o suficiente para serem representativos de toda a língua.

## 4 Por que criar corpora para línguas com poucos recursos?

Dadas as peculiaridades na composição de *corpora* de línguas com poucos recursos descritos na seção anterior, torna-se inevitável questionar se eles realmente valem a pena e que contribuições podem trazer. Para começar a reflexão, é interessante observar a seguinte afirmação de McEnery e Ostler (2000, p.403), "se a linguística de *corpus* é uma abordagem útil na linguística, então ela deveria ser aplicada a todas as línguas". A partir desse imperativo, vale pensar que a Linguística de *Corpus* é uma área multidisciplinar de interesse não só de linguistas, mas também de cientistas computacionais, tecnólogos. E que há pelos menos dois usos possíveis para um *corpus*, como de fonte de dados e como *testbed* para sistemas. Isso significa que quanto mais *corpora,* maiores são as fontes de dados para análises linguísticas e para profissionais de PLN. Para o linguista, o acesso a dados de línguas pouco estudadas pode significar mais argumentos a favor de uma hipótese ou colocar na berlinda teorias que não dão conta de explicar fenômenos raros. Para a área de PLN, o acesso a esses dados pode se apresentar como um desafio e um motivo para a criação de novas ferramentas de processamento, já que, nesse caso, o profissional não poderá sempre usar tecnologias já existentes, pensadas a partir de dados de algumas poucas línguas.

Um outro argumento é que a criação de *corpora* permite que fiquem armazenados conhecimentos de todo tipo, desde históricos e procedimentais a processos evolutivos das sociedades. Também nesse sentido, vale lembrar que a maior parte dos *corpora* de línguas com pouco recursos são oriundos de projetos de documentação de línguas ameaçadas, o que significa que podem ser o único registro dessas línguas, o que é de valor inestimável.

E ainda em relação à documentação de línguas ameaçadas, é importante lembrar que dependendo do grau de ameaça, o tempo que o linguista terá para documentar o máximo possível de dados da língua é muito curto, especialmente quando não há descrições sobre a língua. Se houvesse mais ferramentas computacionais que automatizassem a anotação e a tradução dos dados, o linguista ganharia mais tempo para realizar a documentação e interpretação dos dados coletados. Vale salientar que a pesquisa de campo, para documentação linguística, além de demorada, é cara, e o desenvolvimento de tecnologias computacionais além de otimizar o trabalho, também diminuiria os custos desse tipo de trabalho.

As possibilidades de aplicação de tecnologia linguística são ainda um fator a ser considerado. Como comentam McEnery e Ostler (2000), a maior parte dos sistemas de processamento de língua foram desenvolvidos a partir de línguas de comunidades ricas e poderosas (Estados Unidos, Europa, Japão), enquanto os *corpora* de comunidades pobres e pequenas foram designados simplesmente como repositórios culturais. Uma mudança nesse aspecto poderia, como argumentam os autores, aumentar o acesso de comunidades de línguas não-majoritárias a essas tecnologias e ampliar seu conhecimento técnico e letramento nas novas mídias.

Um último ponto que gostaria de destacar é que para além das possibilidades de desenvolvimento das áreas da Linguística e da Ciência da Computação, a constituição de *corpora* pode ser muito relevante também para as comunidades envolvidas, já que a produção de recursos e o interesse pela língua tendem a resultar na promoção desta entre os membros da comunidade. Como já bastante discutido, a melhora no *status* de uma língua é um dos pontos fundamentais para promovê-la e revitalizá-la.

Por fim, diante de todos os argumentos levantados nessa seção, entendo que a constituição de *corpora* para línguas com poucos recursos, ainda que não cumpram todos os critérios propostos pela Linguística de *Corpus*, devem ser realizados, e os resultados devem ser aproveitados de diversas formas, seja gerando novas tecnologias, servindo de suporte empírico para teorias linguísticas ou promovendo a língua na comunidade.

## Considerações Finais

Iniciamos esse artigo falando do desenvolvimento da área da Linguística de *Corpus*, apresentamos as características de *corpora* "prototípicos" e discutimos as dificuldades para se criar *corpus* gerais que possam ser representativos dos usos de toda uma língua ou variedade, especialmente para *corpora* históricos e de línguas com pouca presença digital. Como vimos, os teóricos da área entendem que um *corpus* deve ser amplo, balanceado e diverso, cobrindo a maior variedade possível de usos de uma língua e apresentando uma boa distribuição entre as subamostras, além de terem de ser finitos e em formato que possa ser processado pelo computador. Seriam, no entanto, poucas as línguas com recursos suficientes para resultarem em um *corpus* balanceado e realmente representativo de todos os usos de uma língua.

A partir de dados sobre a distribuição das línguas no mundo, vimos que menos de 30% delas contam com recursos suficientes para gerarem um *corpus* prototípico e, consequentemente, apenas 30% delas podem contar com estudos linguísticos empíricos. E mais ainda, vimos que a maioria

das teorias linguísticas vigentes são resultado do nosso conhecimento de menos de 4% das línguas humanas, já que apenas línguas oficiais de grandes estados-nação contam com apoio governamental para a sua documentação.

Também mencionei o fato de que tem surgido nas últimas décadas uma série de *corpora* menores, coletados em projetos de documentação linguística dada a iminência de extinção da maior parte das línguas no mundo já neste século. Como discutido, esses *corpora* têm características diferentes dos *corpora* prototípicos, o que torna difícil o desenvolvimento de recursos tecnológicos.

Ao final, concluo que mesmo que os *corpora* de línguas com poucos recursos não tenham a mesma amplitude e os mesmos recursos dos ditos prototípicos, eles são extremamente importantes por várias razões. Podem sim ser uma excelente fonte de dados linguísticos, podem permitir o avanço de tecnologias de PLN, ampliar o acesso a tecnologia e ainda promover o *status* de uma língua entre a comunidade.

## REFERÊNCIAS

ADDA, Gilles; ADDA-DECKER, Martine; AMBOUROUE, Odette; BESACIER, Laurent; BLACHON, David *et al*. Innovative technologies for under-resourced language documentation: The BULB Project. In: WORKSHOP CCURL 2016 – COLLABORATION AND COMPUTING FOR UNDER-RESOURCED LANGUAGES – LREC, 2016, Eslovênia, *Proceedings* […] Eslovênia, HAL, 2016. Disponível em: https://hal.archives-ouvertes.fr/hal-01350124. Acesso em: 10 jan. 2020.

ALLEX – African Languages Lexicon. 2006. Disponível em: http://www.edd.uio.no/allex/ Acesso em: 10 jan. 2020.

ALUISIO, Sandra M.; ALMEIDA, Gladis M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Caleidoscópio,* v. 4, n. 3, p. 156-178, 2006.

ARTS, Tressy *et al*. arTenTen: Arabic Corpus and Word Sketches. *Journal of King Saud University – Computer and Information Sciences*. v. 26, n. 4, p. 357-371, 2014.

ASTON, Guy; BURNARD, Lou. *The BNC Handbook: Exploring the British National Corpus with SARA.* Edimburgo: Edinburgh University press, 1998.

BATISTA, Ronaldo de O. Descrição de Línguas indígenas em gramáticas missionárias do Brasil

Colonial. *D.E.L.T.A*, v. 21, n. 1, p. 121-147, 2005.

BERBER SARDINHA, Tony. *Linguística de corpus*. São Paulo: Manole, 2004.

_____. Tamanho de corpus. *The ESP*, São Paulo, v. 23, n. 2, p. 103-122, 2002.

_____. Linguística de Corpus: Histórico e Problemática. *Delta*, São Paulo, v.12, n.2, p. 323-367, 2000.

BRITISH NATIONAL CORPUS (BNC). Disponível em: <https://www.english-corpora.org/bnc/>. Acesso em: 03 jan. 2020.

COCA – Corpus of Contemporary American English. Disponível em: <*https://www.english-corpora.org/coca/*>. Acesso em: 04 jan. 2020.

COX, Christopher. Corpus Linguistics and Language Documentation: Challenges for Collaboration. *In*: NEWMAN, J.; BAAYEN, H.; RICE, S. (eds.) *Corpus-Based Studies in Language Use, Language Learning, and Language Documentation*. Amsterdam: Rodopi, 2011. p. 239-264.

DIRECT. 2005. Disponível em: http://corpusbrasileiro.pucsp.br/cb/Inicial.html. Acesso em: 03 jan. 2020.

DOBES. 2006. Disponível em: <https://dobes.mpi.nl/projects/?lang=pt>. Acesso em: 02 dez. 2019.

ELDP. 2005. Disponível em: https://www.eldp.net. Acesso em: 02 dez. 2019.

EMILLE – Enabling Minority Language Engineering. 2003. Disponível em: https://www.lancaster.ac.uk/fass/projects/corpus/emille/. Acesso em: 10 jan. 2020.

ENCICLOPÉDIA DAS LÍNGUAS NO BRASIL. Disponível em: https://www.labeurb.unicamp.br/elb/indigenas/documentacao_linguas_indigenas.html. Acesso em: 13 jan. 2020.

GALVES, Charlotte C.; SANDALO, Filomena; SENA, Ticiana A. de; VERONESI, Luis. Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*. Campinas, p. 631-648, set./dez. 2017.

GALÚCIO, Ana Vilacy. Gravações e acervo a partir da pesquisa linguística e cultural, como um passo para a revitalização, fortalecimento e resgate cultural. In: MOREIRA, E; BELAS, C.A.;

et al (orgs.). *Propriedade Intelectual e Patrimônio Cultural: proteção do conhecimento e das expressões culturais tradicionais*, Belém do Pará, MPEG, 2004. p. 109-115.

GRIES, Stefan Th. What is Corpus Linguistics? *Language and Linguistics Compass*, v. 3, p. 1-17, 2009.

GRIES, Stefan Th.; BEREZ, Andrea L. Linguistic Annotation in/for Corpus Linguistics. *In*. IDE, N.; PUTEJOVVSKY, J. *Handbook of Linguistic Annotation*. Berlin – New York: Springer, 2015.

HENNOSTE, Tiit *et al*. Structure and usage of the Tartu University Corpus of Written Estonian. *International Journal of Corpus Linguistics*, v. 3, n.2, p. 279-304, 1998.

JOKINEN, Kristiina. Researching Less-Resourced Languages – the DigiSami Corpus. In: ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC). *Proceedings* […] 2018. p. 3382-3386.

KRAUSS, Michael. The World's Languages in Crisis. *Language* , v. 68, p. 4-10, 1992.

LÜDELING, Anke; KYTÖ, Merja. *Corpus Linguistics: An International Handbook*. V. 1. Berlin – New York: Walter de Gruyter, 2008.

MAXWELL, Mike; HUGHES, Baden. Frontiers in Linguistic Annotation for Lower-Density Languages. *In*: WORKSHOP ON FRONTIERS IN LINGUISTICALLY ANNOTATED CORPORA. Sydney. *Proceedings* […]. Sydney, 2006. p. 29-37.

McENERY, Tony; WILSON, Andrew. *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press, 2001.

McENERY, Tony; OSTLER, Nicholas. A New Agenda for Corpus Linguistics – Working with all of the World's Languages. *Literary and Linguistic Computing*, v. 15, n. 4, p. 403-419, 2000.

OLIVEIRA, Lúcia P.; DIAS, Maria Carmelita P. Compilação de corpus: representatividade e o CORPOBRAS. *Calidoscópio*. v. 7, n. 3, 192-198, set/dez, 2009.

OLIVEIRA Jr., Miguel. NURC Digital. Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). *CHIMERA. Romance Corpora and Linguistic Studies*, v. 3, n. 2, p. 149-174. 2016.

OSTLER, Nicholas. Corpora of less studied languages. *In*.: LÜDELING, A.; KYTÖ, M. *Corpus Linguistics: An International Handbook*. Volume 1. Berlin – New York: Walter de Gruyter, 2008. p. 457-483

SANCHEZ, Aquilino. Definicion e historia de los corpus. In: A. SANCHEZ et al (org.). *CUMBRE – Corpus Linguistico de Espanol Contemporaneo*. Madrid: SGEL, 1995. p. 7-24.

SCANNELL, Kevin P. The Crúbadán Project: Corpus Building for Under-Resourced Languages. *Cahiers du Central*, v. 5, n. 1, p. 1-10, 2007.

SIEPMANN, Dirk; BÜRGEL, Christoph. Le corpus de référence du français contemporain (CRFC), 2014. Disponível em: <https://zenodo.org/record/12353?ln=en#.Xqd1Uy_OpZp>. Acesso em: 27 abr. 2020.

SINCLAIR, Jane M. Corpus and text. Basic principles. In: WYNNE (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 2005. p. 1-16.

SZYMANSKI, Terrence D. *Morphological Inference from Bitext for Resource-Poor Languages*. 2011. Tese (Doutorado em Linguística) – Universidade de Michigan, Michigan, 2011.

TYCHO BRAHE. 2008. Disponível em <http://www.tycho.iel.unicamp.br/~tycho/pesquisa/>. Acesso em 10 dez. 2019.

TRASK, Robert L. *Dicionário de Linguagem e Linguística*. São Paulo: Contexto, 2004.

VINOGRADOV, Igor. Linguistic corpora of understudied languages: Do they make sense? *Káñina, Revista Artes y Letras,* Universidad Costa Rica XL, v.1, p. 127-141, 2016.

ZHOU, Qiang; YU, Shiwen. Annotating the Contemporary Chinese Corpus. *International Journal of Corpus Linguistics*, v. 2, n. 2, p. 199-238. 1997.