# *CORPORA* DIGITAIS PARA ESTUDOS SOBRE O PORTUGUÊS BRASILEIRO CONTEMPORÂNEO – DADOS ORAIS DA REGIÃO SUDOESTE DA BAHIA

Elisângela GONÇALVES
Universidade Estadual do Sudoeste da Bahia
elisangellagoncalves@gmail.com

**Resumo:** Grupos de pesquisas de universidades de diferentes partes do mundo têm se dedicado à constituição de *corpora* anotados sincrônicos e diacrônicos. Quanto ao português, europeu e brasileiro, respectivamente, destacam-se projetos, como o *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*, coordenado pela Profa. Ana Maria Martins, do Centro de Linguística da Universidade de Lisboa, o *Corpus Histórico do Português Anotado Tycho Brahe*, desenvolvido junto ao projeto temático *Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística*, coordenado pela Profa. Charlotte Galves, da Universidade Estadual de Campinas, e o projeto *Corpora digitais para a história do Português Brasileiro – Documentos Históricos da Região Sudoeste da Bahia: Aliança PHPB-Tycho Brahe*, coordenado pelo Prof. Jorge Viana Santos, da Universidade Estadual do Sudoeste da Bahia. Nosso projeto pretende contribuir nessa empreitada, por meio da ampliação das fontes computacionalmente disponíveis para a pesquisa em língua portuguesa. Objetivamos (a) deixar à disposição de pesquisadores dados orais do Português Brasileiro Contemporâneo digitalizados, mais especificamente da variedade da região Sudoeste da Bahia; (b) coletar dados orais urbanos e rurais na região Sudoeste da Bahia, seguindo a metodologia da Sociolinguística Variacionista (LABOV, 1972); (d) realizar a transcrição dos dados, seguindo a metodologia do *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*.

**Palavras-chave:** Corpora digitais; Dados orais; Língua Portuguesa; Morfossintaxe; Região Sudoeste da Bahia.

## 1 Caracterização da comunidade de fala

### 1.1 A Região Sudoeste da Bahia – caracterização

A Região Sudoeste da Bahia é considerada um importante centro empreendedor. Compreende 39 municípios (cf. figura 01), a saber, Anagé, Barra do Choça, Belo Campo, Boa Nova, Bom Jesus da Serra, Caatiba, Caetanos, Cândido Sales, Caraíbas, Cravolândia, Encruzilhada, Firmino Alves, Ibicuí, Iguaí, Irajuba, Itambé, Itapetinga, Itaquara, Itarantim, Itiruçu, Itororó, Jaguaquara, Jequié, Lafayette Coutinho, Lagedo do Tabocal, Macarani, Maiquinique, Manoel Vitorino, Maracás, Mirante, Nova Canaã, Planaltino, Planalto, Poções, Potiraguá, Ribeirão do Largo, Santa Inês, Tremedal e Vitória da Conquista, correspondendo a uma área de 42.542,9 km2, equivalente a 7,5% do estado baiano. De acordo com dados do Instituto de Geografia e Estatística (IBGE), no ano de 2007, verificava-se uma população de aproximadamente 1.144.138, isto é, uma porcentagem aproximada de 8,13% da população do Estado da Bahia (RIBEIRO, 2009).

As principais atividades econômicas da região são a pecuária, sobretudo em Itapetinga, a cafeicultura em Vitória da Conquista, a indústria de transformação, o comércio e os serviços, especialmente nessa cidade e em Jequié. Verifica-se, ainda, na região um elevado crescimento na produção de carnes em decorrência do significativo rebanho bovino e da ampliação da avicultura e suinocultura. Ademais, chama a atenção no Sudoeste da Bahia a produção de leite na bacia do rio Pardo, café em Vitória da

Conquista, Barra do Choça e Planalto, e hortifrutícolas em Jaguaquara e seu entorno. (IBGE, 2013)

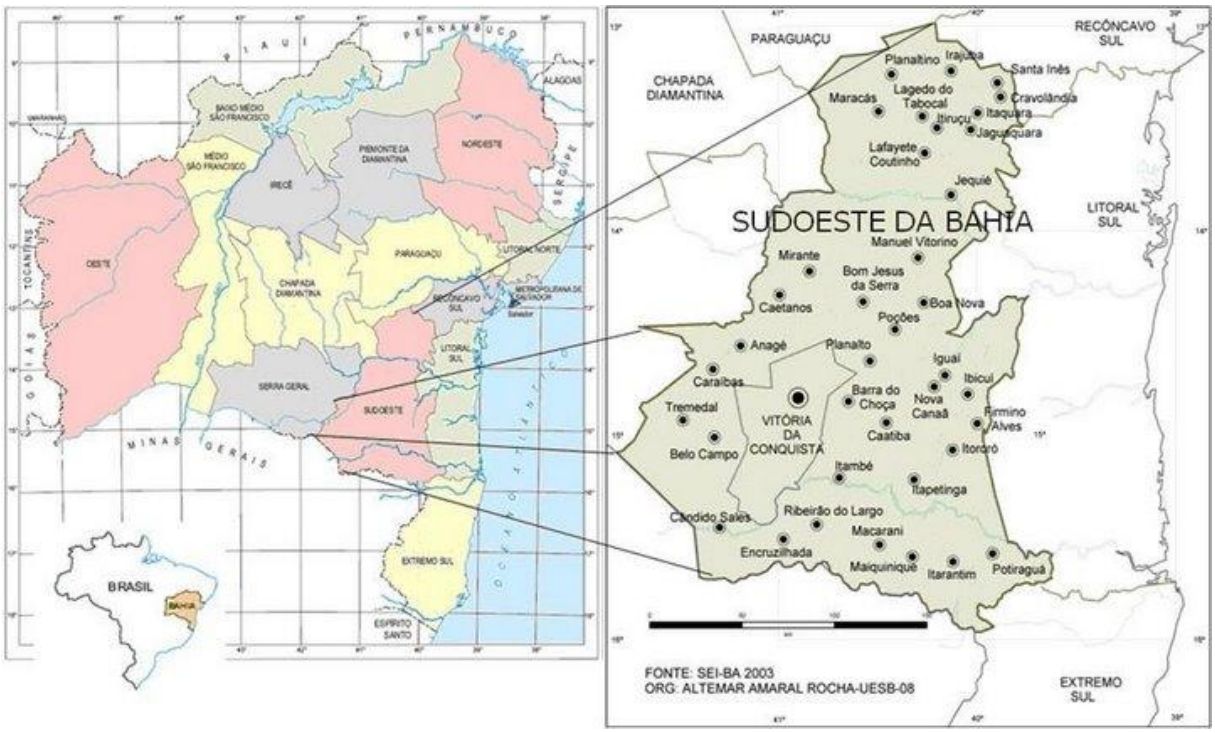
A Região Sudoeste consiste em uma das 15 regiões econômicas do Estado da Bahia propostas pela Superintendência de Estudos Econômicos e Sociais da Bahia – SEI-BA, a partir das décadas de 1980 (final) e de 1990. Consideram-se regiões econômicas e de influência urbana aquelas "destinadas não só a fixação de unidades públicas regionais como de instituições de pesquisa e outros negócios privados". (SANTOS, 2008, p. 37, grifos do autor)

De acordo com Santos,

> Sem sombra de dúvida, esta regionalização que denomina a região em torno de Vitória da Conquista de Região Sudoeste da Bahia é a de maior repercussão entre a sociedade regional. Fica evidente, por exemplo, no emprego da terminologia por instituições públicas e privadas como a Universidade Estadual do Sudoeste da Bahia - UESB; TV Sudoeste da Bahia [...] (SANTOS, 2008, p.37, grifos do autor)

Vitória da Conquista, terceira maior cidade baiana, possui uma população de 336.987 habitantes, distribuídos na sede da cidade e nos onze distritos rurais (conforme dados do IBGE 2013). Encontra-se em uma altitude aproximada de 923 metros. Seu clima se caracteriza como subúmido e seco. Sua temperatura média de 20 graus acarreta um inverno frio e um verão ameno; também é característica da cidade uma leve queda de temperatura no fim de tarde. (http://sevaling2013.com/localizacao/)

Vitória da Conquista constitui um grande centro regional, industrial, comercial e de serviços, pois concentra a maior parte dos investimentos gerados na região. Isso é favorecido por sua localização estratégica ao longo da BR-116 "por onde trafega grande parte de mercadorias que circulam entre o Sudeste e o Nordeste do Brasil". (IBGE, 2013) É nessa cidade que se encontra a sede da Universidade Estadual do Sudoeste da Bahia, com *campi* também em Jequié e Itapetinga. Ainda, no que se refere ao polo educacional, a região conta com um *campus* da Universidade Federal da Bahia (UFBA) na cidade de Vitória da Conquista, bem como com várias faculdades particulares, a exemplo da Faculdade Independente do Nordeste (FAINOR), da Faculdade de Tecnologia e Ciências (FTC), do Instituto Superior Juvêncio Terra e do Instituto Federal da Bahia (IFBA). Assim, concentra aproximadamente 80 municípios da região e do norte de Minas Gerais, o que a faz um importante centro educacional, comercial e social (SEVALING, 2013).




**Mapa Geográfico da Região Sudoeste da Bahia**
**Fonte: tayronefelix.blogspot.com.br**

A Região Sudoeste também se destaca como centro cultural da Bahia, tendo registrados em sua história importantes nomes como o do educador Anísio Teixeira, nascido em Caetité, de Glauber Rocha, cineasta que marcou o cinema novo, e Elomar Figuiera, cantor e compositor, nascidos em Vitória da Conquista, de Waly Salomão, poeta brasileiro nascido em Jequié.

### 1.1 Região Sudoeste da Bahia – características linguísticas

Pesquisadores da Universidade do Sudoeste da Bahia, da Área de Língua Portuguesa e Linguística, têm desenvolvido pesquisas com *corpora* orais ou escritos da Região Sudoeste, sobretudo com dados de falantes do município de Vitória da Conquista, em diversas áreas de pesquisa: (a) fonética/fonologia, em que são desenvolvidos estudos sobre a caracterização acústica das exclamativas e interrogativas do dialeto de Vitória da Conquista-BA; sobre a variação do /S/ em posição de trava silábica e o processo de debucalização no dialeto de Vitória da Conquista-BA; (b) morfologia/sintaxe, seja sob a perspectiva gerativista – estudo de nós e a gente, investigando-se se se trata de variação ou de mudança – seja sob a perspectiva da sociolinguística variacionista – análise da concordância verbal em P6 no português popular urbano de Vitória da Conquista-BA – seja sob o viés da aquisição da linguagem, voltando-se para a aquisição das estratégias de relativização em PB, com base em dados de fala de crianças conquistenses, por exemplo.

Consideramos de grande importância a introdução de dados de falantes de diferentes partes do país na análise linguística, pois, somente desse modo, será possível caracterizar, de fato, os diversos falares do Brasil, que não se constitui em um país linguisticamente homogêneo. Nesse sentido, alguns estudos desenvolvidos a partir de dados da Região Sudoeste têm trazido novos questionamentos sobre o que têm sido proposto a respeito de alguns fenômenos linguísticos do Português Brasileiro Contemporâneo, enquanto outros corroboram as hipóteses apresentadas. Entre aqueles cujos resultados apontam "desafios" ao que tem sido posto na literatura linguística corrente, destaco três:

(a) Silva (2005) analisa a concordância verbal na terceira pessoa do plural, em três comunidades linguísticas do interior (da região Sudoeste) do estado da Bahia – duas rurais (Cinzentos e Morrinhos) e uma urbana (Poções). Os resultados desse pesquisador vai de encontro às pesquisas realizadas em outras regiões do país sobre o a concordância de terceira pessoa do plural, visto que "a concordância verbal se apresenta como uma regra variável, demonstrando, contudo, fortes tendências à aquisição de padrões que conduzam à aplicação da regra de concordância, aproximando-se cada vez mais da norma culta urbana" (SILVA, 2005, resumo). Essa tendência vem condicionada (a) pelo nível de escolarização, que, ainda que precário, pode levar o falante a manter contato com outros grupos sociais, assim como a empregar um novo padrão mais próximo do português culto; (b) pelo contato deste com os meios de comunicação, sobretudo televisão e rádio, elementos difusores da norma de prestígio. Ainda, segundo o autor, "o estágio de urbanização pode favorecer o processo de aquisição de um padrão linguístico tendente a apresentar progressivamente a ocorrência da concordância entre o sujeito e o predicado" (Idem).

(b) Gonçalves (2004), num estudo sobre o emprego do sujeito nulo no Português Brasileiro Contemporâneo (PBC), verificou, a partir da quantificação e interpretação dos dados, que o português brasileiro falado na comunidade de Vitória da Conquista não demonstra estar se afastando do grupo de línguas de sujeito nulo, como propõem trabalhos como o de Duarte (1995). A autora assume que *pro*, no PBC, não é mais licenciado pela flexão verbal, visto que essa variedade do português não mais apresenta um sistema verbal com flexão rica em decorrência da reestruturação do seu sistema pronominal, mas de duas outras maneiras: (i) por meio da correferência com um tópico discursivo, em orações matrizes. Nesse sentido, o PBC estaria se assemelhando a uma língua como o chinês, direcionada para o discurso, e que apresenta sujeitos nulos a despeito da ausência de uma flexão rica para licenciá-los, conforme afirma Negrão (1990); (ii) por meio da correferência com um sujeito da oração matriz, em orações encaixadas.

(c) Gonçalves (2013) descobriu em dados de Vitória da Conquista ***construções existenciais com o verbo ser***, **cuja realização fora verificada somente até a segunda metade do século XVI** (RIBEIRO, 1996, p. 377; MATTOS E SILVA, 1994, p. 11), não consistindo, entretanto, em uma extensão ou resíduo das construções existenciais com *ser* do Português Medieval, segundo a autora. Fica em aberto para pesquisas futuras a questão de saber quais fatores desencadearam a emergência dessas construções.

Cremos que esses trabalhos contribuirão para os estudos linguísticos desenvolvidos acerca da caracterização do português brasileiro. Tal caracterização só poderá ser concluída a partir de quando pesquisadores de diferentes partes do Brasil procederem a análises de diferentes fenômenos linguísticos.

## 2 Delimitação do objeto de estudo e justificativa

Grupos de pesquisas de universidades de diferentes partes do mundo têm se dedicado à constituição/exploração de *corpora* anotados sincrônicos e diacrônicos. No que diz respeito ao português, europeu e brasileiro, respectivamente, destacam-se projetos, como o *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*, coordenado pela Profa. Ana Maria Martins, do Centro de Linguística da Universidade de Lisboa, o *Corpus Histórico do Português Anotado Tycho Brahe*, desenvolvido junto ao projeto temático *Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística*, coordenado pela Profa. Charlotte Galves, da Universidade Estadual de Campinas, e o projeto *Corpora digitais para a história do*

*Português Brasileiro – Documentos Históricos da Região Sudoeste da Bahia: Aliança PHPB-Tycho Brahe*, coordenado pelo Prof. Jorge Viana Santos, da Universidade Estadual do Sudoeste da Bahia. No decorrer deste projeto, abordaremos com mais detalhes o CORDIAL-SIN, que, por sua vez, segue a metodologia de notação sintática do *Corpus* Tycho Brahe.

O presente projeto pretende contribuir nessa empreitada, por meio da ampliação das fontes computacionalmente disponíveis para a pesquisa em língua portuguesa. Através da disponibilização de dados em meios digitais (na internet), pretendemos contribuir com pesquisadores que desenvolvem pesquisas linguísticas sobre fenômenos do português, sejam sincrônicas sejam diacrônicas. Esses fatores justificam a sua consecução.

Nesta pesquisa, lançaremos mão da notação sintática empregada no *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)* e desenvolveremos buscas utilizando um programa específico para a busca em textos anotados sintaticamente, denominado *Corpus Search*.

## 2.1 AHDig: Associação das Humanidades Digitais

Este projeto se encontra ligado à **Associação das Humanidades Digitais,** uma rede de "pesquisadores de diferentes áreas das humanidades e das ciências da informação e da computação, em universidades brasileiras, portuguesas e outras, envolvidos em diversos grupos de pesquisas, projetos e iniciativas ligados às Humanidades Digitais [...] unidos pela língua portuguesa e pela inclusão da perspectiva digital em seus horizontes de pesquisa" (ASSOCIAÇÃO DAS HUMANIDADES DIGITAIS, 2013). Essa Associação foi criada por pesquisadores brasileiros e portugueses em 25 de outubro de 2013, visando, sobretudo, ao fortalecimento das "iniciativas em Humanidades Digitais já ativas no universo dos falantes do português, e promover novas iniciativas nesse campo entre eles" (Ibid.). Nessa rede, estudiosos de quaisquer países interessados em desenvolver pesquisas sobre o português terão a oportunidade de discutir e trocar experiências.

A ideia de formar tal associação surgiu de debates travados durante as reuniões do *Grupo de Pesquisas Humanidades Digitais*, na Universidade de São Paulo, ganhando forma no decorrer da organização do *Dia das Humanidades Digitais em Português e Espanhol*, consolidando-se na *I Oficina sobre Construção e Uso de Grandes Corpora* (12 de setembro de 2013), e no *I Seminário Internacional em Humanidades Digitais no Brasil*, realizado entre 22 e 25 de outubro de 2013, quando foi oficialmente fundada a **AHDig, Associação das Humanidades Digitais**. Atualmente, compõe a AHDig um grupo de vinte e seis pesquisadores.

Segundo Paixão:

> Inicialmente, funcionaremos como uma rede virtual, por meio deste espaço e de uma lista de discussões a ser implementada em breve. Até o início de 2014, a Comissão de Fundação tem como tarefa apresentar uma proposta para o funcionamento consolidado da Associação, deliberando seus mecanismos futuros de filiação e as atividades que serão abrigadas na entidade.

Atualmente a Rede AHDig conta com os seguintes projetos de pesquisa, sendo que existem outros (como nosso projeto) que não constam na página da associação por não disporem, ainda, de um sítio eletrônico:

- **Atlas das Paisagens Literárias de Portugal Continental,**
  *Universidade Nova de Lisboa*:
  http://paisagensliterarias.ielt.org/

- **Atlas, Cartografia Histórica,**
  *Universidade Nova de Lisboa*:
  http://atlas.fcsh.unl.pt/

- **Brasiliana USP**,
  *Universidade de São Paulo*:
  http://www.brasiliana.usp.br

- **Caminhos do Romance**,
  *Universidade Estadual de Campinas*:
  http://www.caminhosdoromance.iel.unicamp.br/

- **Circulação Transatlântica dos Impressos – a globalização da cultura no século XIX**,
  *Universidade Estadual de Campinas*,
  http://www.circulacaodosimpressos.iel.unicamp.br/

- **Corpus Anotado do Português Histórico Tycho Brahe,**
  *Universidade Estadual de Campinas*:
  http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html

- **Corpus Eletrônico de Documentos Históricos do Sertão (CEDOHS),**
  *Universidade Estadual de Feira de Santana*:
  http://www2.uefs.br/cedohs/

- **Edições Digitais para a História da Língua Portuguesa (EDHILP),**
  *Universidade de Évora*:
  http://host.di.uevora.pt/edhilp/

- **Edição dos Panfletos de Eulálio Motta**,
  *Universidade Federal da Bahia*:
  http://www.eulaliomotta.com.br/

- **eDictor: ferramenta para edição filológica eletrônica**,
  *Universidade de São Paulo e Universidade Estadual de Campinas*
  (desenvolvimento): http://manualedictor.wordpress.com/

- **Post Scriptum, Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna (P.S.),**
  *Universidade de Lisboa*:
  http://ps.clul.ul.pt/index.php

e com os seguintes grupos de pesquisa:

- **Grupo de Pesquisas História, Mapas e Computadores (Hímaco),**
  *Universidade Federal de São Paulo / Arquivo do Estado de São Paulo*:
  http://www.arquivoestado.sp.gov.br/himaco/

- **Grupo de Pesquisas Humanidades Digitais,**
  *Universidade de São Paulo*:
  http://humanidadesdigitais.org/

- **Grupo de Pesquisas Linguagens e Educação em Rede (LER),**
  *Universidade Federal do Ceará*:
  http://www2.virtual.ufc.br/ler/

## 3 Objetivos, hipóteses e metas

Objetivamos neste estudo:

- Deixar à disposição de pesquisadores dados orais (rurais e urbanos) do Português Brasileiro Contemporâneo digitalizados, mais especificamente da variedade da região Sudoeste da Bahia, a fim de que os mesmos possam realizar estudos em diferentes áreas da linguística;

- Somar-se a projetos que têm por finalidade a elaboração e implementação de *corpora* anotados em meio digital;

- Coletar dados orais urbanos e rurais na região Sudoeste da Bahia, seguindo a metodologia da Sociolinguística Variacionista (LABOV, 1972);

- Realizar a transcrição dos dados, seguindo a metodologia do *Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*;

- Orientar discentes de Iniciação Científica e de mestrado, posteriormente.

## 4 Fundamentação teórica

### 4.1 Teoria Gerativa

As pesquisas que serão realizadas no âmbito desse projeto e que se utilizarão desse banco de dados tomarão como base a Teoria Gerativa numa perspectiva sincrônica, embora tais pesquisas sincrônicas fomentadas pelo banco de dados orais do Sudoeste da Bahia possam ser utilizadas para um estudo diacrônico contemplando outros *corpora*. Nesse sentido, este projeto mantém parcerias com o projeto temático: "Sintaxe diacrônica em corpus eletrônico: do português pré-classico às variantes modernas", coordenado pela Profa. Dra. Cristiane Namiuti Temponi (FAPESB/UESB) e com os projetos "Memória Conquistense: Recuperação de Documentos oitocentistas de Vitória da Conquista" (sem ônus, coordenado pelos professores doutores Jorve Viana Santos e Cristiane Namiuti Temponi, 2009) e o projeto financiado pela FAPESB "Corpora Digitais para a história do português Brasileiro: Aliança PHPB Tycho Brahe (SANTOS & NAMIUTI, 2010).

A questão central da Teoria Gerativa, principiada pelo norte-americano Noam Chomsky a partir dos anos 1950, é a aquisição da linguagem.

O gerativismo se inscreve na corrente naturalista dos estudos sobre a linguagem e a formação humana. Sua natureza é *mentalista*, já que concebe a língua como um sistema de regras e princípios arraigados na mente humana, e as línguas naturais como adquiridas e faladas espontaneamente apenas pelos membros da espécie humana (cf. RAPOSO, 1992). A Teoria Gerativa, todavia, não nega o papel do ambiente na aquisição da linguagem.

Lightfoot (1999, p. 52) afirma: "Language emerges through an interaction between our genetic inheritance and the linguistic environment". As gramáticas consistem em entidades que surgem nas mentes dos indivíduos ao serem expostos à experiência, quando crianças.

Clark & Roberts (1993) afirmam que o objetivo da criança é adequar o conjunto da experiência linguística primária a gramáticas hipotéticas. Nesse momento da trajetória dos estudos gerativistas, a mudança linguística no plano dinâmico e processual, começou a ser vista como teoricamente relevante, tendo em vista que foi percebida a importância dos processos de mudança para o entendimento da *aquisição da linguagem*, ou seja, percebeu-se a contribuição a ser dada pelos estudos da mudança "para a compreensão do objeto teórico faculdade da linguagem – mais especificamente, por elucidar a relação entre o ambiente linguístico e a *gramática universal* no processo de emergência das gramáticas nos falantes" (PAIXÃO DE SOUSA, 2006, p. 22). Houve uma alteração de perspectiva das pesquisas diacrônicas no quadro gerativista: passou-se da focalização "da comparação de sincronias para a própria dinâmica da mudança" (Ibid.). Seus estudos têm focalizado as abordagens quantitativas, valendo-se dos estudos estatísticos e populacionais do sócio-variacionismo, sobretudo a partir da década de 1990.

A dimensão dinâmica dos estudos diacrônicos no gerativismo consiste na passagem de uma gramática antiga a uma nova, ou seja, a **Mudança de Gramáticas**.

Conforme Paixão de Sousa (2006, p. 21), a *mudança* pode consistir num objeto fundamental de reflexão mesmo para a Teoria Gerativa, para a qual a instabilidade, a impermanência, a heterogeneidade não constituem o foco principal de estudo.

Nesse sentido, desenvolveram-se muitos estudos diacrônicos gerativistas, sobretudo no início dos anos 1990, como os de Lightfoot (1999), Clark & Roberts (1993).

Mattos e Silva (1994), por sua vez, atribui ao desenvolvimento da Sociolinguística Variacionista (WEINREICH, LABOV e HERZOG, 1968) e do modelo de Princípios e Parâmetros (CHOMSKY, 1981) o retorno dos estudos histórico-diacrônicos no Brasil, abandonado entre os anos de 1960 e 1980.

O modelo de Princípios e Parâmetros e, em alguns estudos mais recentes, o Programa Minimalista (seu desdobramento/afunilamento), é que tem servido como base para os estudos comparativos dos últimos vinte anos.

## 5 Materiais e Métodos

A metodologia a ser por nós utilizada na coleta e seleção dos informantes será pautada na Sociolinguística Variacionista (LABOV, 1972).

Como se trata de um projeto em que serão envolvidos seres humanos, por questões éticas, seguiremos estritamente as normas estabelecidas pelo Comitê de Ética em Pesquisa da Universidade Estadual do Sudoeste da Bahia, ressaltando que o mesmo está sendo submetido à Plataforma Brasil. Como parte das exigências para a aprovação deste projeto, elaboramos um Termo de Consentimento Livre e Esclarecido – TCLE (modelo em Anexos), cujo objetivo é o de manter os informantes (entrevistados) cientes sobre fatores relevantes da pesquisa, como o seu objetivo, as justificativas para a sua consecução, entre outros.

### 5.1 População e amostragem

A princípio, ocupar-nos-emos da coleta de dados da zona urbana do município de Vitória da Conquista cujo número de informantes a compor a amostragem da população será de 72, distribuídos da seguinte forma:

1- Gênero:

Masculino: 36 informantes
Feminino: 36 informantes

2- Faixa etária:
15 a 25 anos: 24 informantes
35 a 45 anos: 24 informantes
Mais de 50 anos: 24 informantes

3 - Nível de escolarização:
Fundamental: 24 informantes
Médio: 24 informantes
Superior: 24 informantes

**5.2 Seleção dos informantes**

Objetivando selecionar os informantes que constituirão a amostragem da população, será utilizada a técnica de amostra aleatória por área.

Inicialmente, serão sorteados cinco (5) bairros, e, em seguida, duas ruas por bairro, onde será aplicado um total de quinhentos (500) questionários, em Vitória da Conquista, com vistas a selecionar os prováveis informantes. A partir da lista nominal dos informantes selecionados pelo primeiro questionário, serão escolhidos, através da técnica de amostra aleatória simples, os setenta e dois (72) que comporão a amostragem final.

Na escolha dos informantes, serão observados os seguintes requisitos:

a) ser natural de Vitória da Conquista ou morar neste município desde os cinco anos de idade.
b) nunca ter passado mais do que dois anos consecutivos fora desta cidade.

**5.3 Instrumentos de pesquisa e coleta de dados**

Para obter os dados que comporão o *corpus* da pesquisa, será, inicialmente, aplicado um questionário (anexo I), que possibilitará a seleção dos informantes, e, em seguida, após definição da amostragem, aplicar-se-á uma ficha social (anexo II), com o objetivo de se manter o primeiro contato com os informantes, visando à minimização do ***paradoxo do observador,*** como salienta Labov (1972, p. 209).

Vencidas essas duas etapas, na seguinte, será feita uma entrevista com cada informante, que constará de questões voltadas para o cotidiano, levando-se em conta as especificidades de cada informante, detectadas a partir da ficha social anteriormente aplicada. Cada entrevista deverá ter a duração de aproximadamente 60 minutos.

Os dados serão, posteriormente, transcritos conforme as normas do CORDIAL-SIN.

**5.4 Modelo de transcrição de dados adotado pelo projeto *Corpus Dialectal para o Estudo da Sintaxe* (CORDIAL-SIN)**

O projeto *Corpus Dialectal para o Estudo da Sintaxe* (CORDIAL-SIN) realiza análise de fenômenos sintáticos do português europeu, por meio de uma metodologia de constituição/exploração de um *corpus* anotado, cuja extensão é de 600.000 palavras e tem como base a teoria de Princípios e Parâmetros (CHOMSKY, 1995). Esse projeto conta com o financiamento da Fundação para a Ciência e a Tecnologia – FCT.

Estão entre os objetivos do Cordial-SIN: (a) desenvolver um estudo comparado da sintaxe dos dialetos do português europeu; (b) reforçar a cooperação com projetos internacionais de sintaxe dialetal; (c) elaborar e disponibilizar *corpus* digital, a fim de atingir os objetivos anteriores.

O Cordial-SIN faz parte de "um conjunto geograficamente representativo de excertos de discurso livre e semi-dirigido seleccionados a partir das gravações efectuadas" nos projetos Atlas Linguístico e Etnográfico de Portugal e Galiza (ALEPG); Atlas Linguístico do Litoral Português (ALLP); Atlas Linguístico e Etnográfico dos Açores (ALEAc); Fronteira Dialectal do Barlavento Algarvio (BA) (DESCRIÇÃO DO PROJECTO CORPUS DIALETAL PARA O ESTUDO DA SINTAXE, 2013, p. 1). Tais projetos estão vinculados ao Grupo de Dialectologia do Centro de Linguística da Universidade de Lisboa (CLUL), que nas últimas três décadas constituiu um arquivo com aproximadamente 4.500 horas de gravações, realizadas em mais de 200 localidades de Portugal.

Estão disponibilizados ao consultor quatro formatos do CORDIAL-SIN:

(a) transcrição conservadora: em que são registradas informações sobre aspectos da produção, como pausas, sobreposições de produção, hesitações, abandono de fragmentos frásicos, reformulações, repetições, formas truncadas, variantes fonéticas e morfofonológicas, entre outras;

(b) transcrição ortográfica normalizada: conta somente com transcrição ortográfica, sendo obtida por meio da exclusão das informações sobre aspectos da produção e consiste no suporte da anotação. Contempla tanto frases completas quanto frases inacabadas – sintaticamente analisáveis e anotáveis –, chamadas fragmentos frásicos;

(c) texto com anotação morfossintática (anotação por palavra): é automaticamente implementada; consiste em uma revisão/ampliação do sistema desenvolvido pelo grupo do projeto Tycho Brahe. A proximidade entre os sistemas de anotação morfossintática do CORDIAL-SIN e do projeto Tycho Brahe permite a utilização do etiquetador automático, de base probabilística, desenvolvido por Marcelo Finger (e melhorado por Fabio Natanael Kepler e Marcelo Finger) no âmbito do Tycho Brahe. Neste sistema de anotação, são combinadas etiquetas categoriais e subetiquetas, sobretudo flexionais, o que possibilita uma anotação apurada das unidades lexicais do *corpus*;

(d) texto com anotação sintática (anotação por frase): toma como base as orientações do Penn-Helsinki Parsed *Corpus* of Middle English; assim como "opera sobre dados etiquetados morfossintacticamente; marca fronteiras de constituintes, dependências sintagmáticas e oracionais, tipos de frases, relações gramaticais e certas relações transformacionais [...]" Ainda, "define configurações que podem ser pesquisadas sistemática e exaustivamente [...] compatíveis com o programa Corpus Search 2, da autoria de Beth Randall (open source software, Sourceforge)" (DESCRIÇÃO DO PROJECTO CORPUS DIALETAL PARA O ESTUDO DA SINTAXE, 2013, p. 2).

Nesta pesquisa, utilizaremos a versão conservadora, sobre a qual discorremos um pouco mais, voltando-nos especificamente para as normas de transcrição.

O início de cada transcrição é marcado por um cabeçalho composto pelo código de identificação do ficheiro (proveniência geográfica do texto, caracterização dos informantes, identidade dos inquiridores e dos transcritores, identificação do inquérito no arquivo sonoro do Grupo de Estudos de Dialectologia do CLUL, localização do texto transcrito no arquivo sonoro do CORDIAL-SIN, data da transcrição). No entanto, nem sempre é possível preencher completamente os campos relativos à caracterização dos informantes, devido à falta de informações na fonte. De modo a preservar-se a identidade destes, dá-se a ele um nome próprio fictício. Quanto ao assunto, existe uma lista que pode ser utilizada nas entrevistas; tal lista é estabelecida com base no índice do questionário linguístico do projeto Atlas Linguístico-Etnográfico de Portugal e da Galiza, desenvolvido pelo Grupo de Estudos de

Dialectologia do CLUL. Todavia, em se tratando de fragmentos curtos ou de caráter geral, poderá usar-se a expressão "não aplicável" no campo *assunto*. Segue abaixo o formato do cabeçalho.

| | |
|---|---|
| Código de identificação do ficheiro: | |
| Localidade:<br>Distrito: | Concelho:<br>Data: |
| Informante1:<br>Idade: | Sexo:<br>Escolaridade: |
| Informante2:<br>Idade: | Sexo:<br>Escolaridade: |
| Informante3:<br>Idade: | Sexo:<br>Escolaridade: |
| Fonte:<br>Inquiridor1:<br>Cassete nº: lado: min: | Inquiridor2: |
| Assunto: | |
| Tipo de transcrição:<br>Autor da primeira transcrição:<br>Autor da revisão final:<br>CD nº: faixa: | Data da primeira transcrição:<br>Data da revisão final: |

No que se refere aos procedimentos de identificação dos interlocutores, cada tomada de palavra inicia um parágrafo e é introduzida pelas abreviaturas "INQ" ou "INF", em se tratando, respectivamente, da fala do inquiridor e do informante, sem que haja nenhum sinal de pontuação entre elas e o início dos enunciados. Quando houver mais de um inquiridor e/ou informante, tais abreviaturas serão seguidas de um algarismo indicando "a ordem de intervenção no diálogo por parte de cada locutor", conforme exemplos em (1) a seguir. (NORMAS DE TRANSCRIÇÃO DO PROJECTO CORPUS DIALETAL PARA O ESTUDO DA SINTAXE, 2013, p. 4).

(1) *INQ1 Pois, fica, entra nos dois, pois…*
INF2 O senhor já viu, o senhor já viu o que isto mata?
*INQ1 Como é que chama a este?*
INF2 Faneca.
INF1 É fodãozinho.
*INQ2 A faneca…*
(VPA02-C)

A transcrição é feita de acordo com a ortografia oficial portuguesa, com algumas ressalvas que não discutiremos aqui.

Numa etapa futura do projeto, pretendemos realizar a anotação morfossintática (palavra por palavra), inspirado no sistema de anotação do projeto Tycho Brahe, conforme mencionado acima, bem como a anotação sintática.

## 6 Considerações Finais

A consecução deste projeto será importante, visto que, ao disponibilizarmos dados digitais para consulta por pesquiadores em linguística, possibilitaremos aos mesmos a descrição e análise de uma extensa quantidade de dados com precisão e agilidade por meio do emprego de ferramentas automáticas. Ademais, uma vez que, por meio da pesquisa linguística, é possível conhecer-se a língua (em nosso caso, a língua portuguesa) como

realmente é produzida por seus falantes, indo além do que é documentado e registrado pela gramática normativa (tradicional) e das regras por ela ditadas, pesquisadores da língua portuguesa se beneficiarão à medida que disporão de dados de mais uma comunidade linguística desse imenso país que é o Brasil.

**7 Referências**

ASSOCIAÇÃO DAS HUMANIDADES DIGITAIS, 2013. Disponível em: < http://ahdig.org/2013/11/05/breve-historico-da-ahdig/>. Acesso em: 30 nov. 2013.

CHOMSKY, N. Principles and Parameters in syntactic theory. In: HORNSTEIN, N.; LIGHFOOT, D. (Ed.). **Explanations in Linguistics**. New York: Longman, 1981.

_____. **The Minimalist Program**. Cambridge, Massachussets: MIT Press, 1995.

CLARK, R.; ROBERTS, I. A computational model of language learnability and language change. **D.E.L.T.A.**, n. 8, 1993, p. 53-103.

DESCRIÇÃO do Projecto Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN). In: *Centro de Linguística da Universidade de Lisboa*, 2013. Disponível em: < http://www.clul.ul.pt/pt/recursos/225-description-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects?format=pdf>. Acesso em: 10 jun. 2013.

GONÇALVES, E. **O sujeito nulo na comunidade linguística de Vitória da Conquista-BA.** 2004. 200 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2004.

_____. **SER OU NÃO SER:** eis a questão - Construções Existenciais com o Verbo Ser no Português Brasileiro Contemporâneo. 2013. 162 f. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2013.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2013. Disponível em: < www.ibge.gov.br>. Acesso em: 20 out. 2013.

I SEMINÁRIO DE VARIAÇÃO E MUDANÇA LINGUÍSTICA NO SUDOESTE DA BAHIA, 2013. Disponível em < http://sevaling2013.com/>. Acesso em: 30 nov. 2013.

LABOV, W. **Sociolinguistic patterns.** Philadelphia: University of Pennsylvania Press, 1972.

LIGHTFOOT, D. **The development of language:** Acquisition, change, and evolution. Malden: Blackwell/Maryland lectures in language and cognition, 1999.

MAPA DA REGIÃO SUDOESTE DA BAHIA, 2009. Disponível em: <http://tayronefelix.blogspot.com.br/2009/01/mapa-regiao-sudoeste-da-bahia.html>. Acesso em: 20 out. 2013.

MATTOS e SILVA, R. V. **O Português Arcaico**: Morfologia e Sintaxe. São Paulo: Contexto, 1994.

NORMAS de Transcrição do Projecto Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN). In: *Centro de Linguística da Universidade de Lisboa*, 2013. Disponível em: <http://www.clul.ul.pt/pt/recursos/225-description-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects?format=pdf>. Acesso em: 10 jun 2013.

PAIXÃO DE SOUSA, M. C. Lingüística Histórica. In: **II Escola de Verão de Lingüística Formal da América do Sul**, 2005, Campinas: Unicamp. Disponível em: <http:////www.ime.usp.br/~tycho/participants/psousa>. Acesso em: 20 nov. 2006.

RAPOSO, E. P. **Teoria da Gramática**: a faculdade da linguagem. Lisboa: Caminho, 1992.

RIBEIRO, I. A formação dos tempos compostos: a evolução histórica das formas *ter*, *haver* e *ser*, In: ROBERTS I.; KATO, M. A. (Org.). **Português brasileiro**: uma viagem diacrônica. 2. ed. Campinas, Editora UNICAMP, 1996, p. 343-386.

RIBEIRO, V. B. **Estudo Regional do Sudoeste da Bahia.** Disponível em: <http://veranilzabr.blogspot.com.br/2009/01/apresentando-regio-sudoeste-da-bahia.html>. Acesso em: 20 jun. 2013.

SANTOS, V. C. C. **Estado da Bahia:** Novas e velhas regionalizações (identificação, listagem e ilustração). Vitória da Conquista: texto digitado, 2008.

SILVA, J. A. A. da. **A concordância verbal de terceira pessoa do plural no português popular do Brasil:** um panorama sociolingüístico de três comunidades do interior do Estado da Bahia. 2005. 323 f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2005.

SUPERINTENDÊNCIA ESTUDOS ECONÔMICOS E SOCIAIS DA BAHIA, 2013. Disponível em: <www.sei.ba.gov.br>. Acesso em: 20 out. 2013.

WEINREICH, U., LABOV, W. e HERZOG, M. Empirical foundations for a theory of language change. In: **Directions for historical linguistics**. Austin, University of Texas Press, 1968.

**ANEXO I**
**QUESTIONÁRIO**

1 – Nome: _____

2 – Endereço: _____
_____

3 – Gênero:
( ) masculino
( ) feminino

4 – Local de nascimento: _____

5 – Idade: _____

6 – Sempre residiu em _____?
( ) Sim
( ) Não

6.1 – Se não, por quanto tempo residiu fora de _____?
( ) menos de dois anos consecutivos
( ) mais de dois anos consecutivos

7 – Anos de escolarização:
( ) nenhum
( ) 1 a 4 anos
( ) 5 a 8 anos
( ) 9 a 11 anos
( ) 12 anos em diante

8 – Caso venha a ser selecionado, dispõe-se a responder algumas perguntas para uma pesquisa que estamos realizando?
( ) Sim
( ) Não

9 – Em caso afirmativo, qual o melhor horário e local?
Horário: _____
Local: _____

**ANEXO II**
**FICHA SOCIAL**

1 – Nome: _____

2 – Endereço residencial: _____
_____

3 – Gênero:
( ) masculino
( ) feminino

4 – Idade: _____

5 – Estado civil:
( ) solteiro
( ) casado
( ) viúvo
( ) separado / divorciado
( ) união sem vínculo oficial

6 – Tipo de moradia:
( ) própria
( ) alugada
( ) cedida
( ) outro – Especificar: _____

7 – Tipo de construção:

( ) taipa
( ) tijolo
( ) madeira
( ) outro – Especificar: _____

8 – Iluminação da casa:
( ) elétrica
( ) querosene
( ) outro – Especificar: _____

9 – Na casa, qual o número de:
( ) salas: ____
( ) quartos: ____
( ) banheiros: ____
( ) cozinhas: ____
( ) terraços: ____
Total: ____

10 – Você possui:
( ) televisão
( ) aparelho de som
( ) aparelho de DVD

( ) computador (notebook, netbook, tablet)
( ) aparelho celular
( ) liquidificador
( ) aspirador de pó
( ) geladeira
( ) máquina de lavar roupa
( ) máquina de lavar louça
( ) forno microondas
( ) bicicleta
( ) motocicleta
( ) automóvel

11 – Ocupação profissional: _____

12 – Horário de trabalho: _____

13 – Renda mensal: _____

14 – Grau de escolarização: _____

15 – Atualmente, a maneira como está vivendo lhe dá:
( ) muita satisfação
( ) pouca satisfação
( ) nenhuma satisfação

16 – Em geral, como acha que vão as coisas atualmente?
( ) muito bem
( ) bem
( ) mais ou menos
( ) mal

17 – Este é um desenho de uma escada:

| |
|---|
| _1_ |
| _2_ |
| _3_ |
| _4_ |
| _5_ |
| _6_ |
| _7_ |
| _8_ |
| _9_ |
| _1 |
| 0 |

Suponha que último degrau da escada (nº 1) represente a melhor vida possível e o degrau nº 10, a pior. Onde você se localizaria? (Marcar com um X o degrau correspondente.)

18 – Você é uma pessoa que:
( ) nunca sai do município onde reside
( ) só sai a negócio

( ) sempre sai para passear

19 – Se sai, passa quanto tempo fora?
( ) menos de um mês
( ) de um a três meses
( ) mais de três meses – Especificar: _____

20 – Se viaja a passeio para outros estados, o que costuma fazer:
( ) ir a cinemas
( ) ir a teatros
( ) ir a estádios de futebol
( ) ir à praia
( ) ir a livrarias
( ) fazer compras
( ) apenas descansar

21 – A cidade em que você mora é:
( ) muito bonita
( ) boa
( ) regular
( ) muito atrasada

22 – Se tivesse que morar em outra cidade, onde moraria? _____