Transcrição automática de entrevistas e anotação Universal Dependencies no Corpus Roda Viva

DOI: http://dx.doi.org/10.21165/el.v54i1.3851

Cláudia Dias de Barros[1]
Oto Araújo Vale[2]
Gabriela Wick-Pedro[3]

Resumo

Neste artigo é apresentada a pesquisa sobre a transcrição automática de quatro entrevistas extraídas do Corpus Roda Viva, que é formado por 713 entrevistas do Programa Roda Viva, da TV Cultura. As entrevistas originais foram transcritas por jornalistas, adquirindo, assim, um *status* de texto escrito, possuindo, ainda, intervenções, como informações enciclopédicas sobre fatos e pessoas citadas. A fim de trabalhar com texto oral, a presente pesquisa realizou um trabalho piloto de transcrição automática de quatro dessas entrevistas, usando a ferramenta Whisper e, posteriormente, as entrevistas foram anotadas automaticamente com a formalização da Universal Dependencies e revisadas manualmente pela ferramenta Arborator Grew ElizIA. Por meio desse trabalho, pôde-se notar as diferenças sintáticas presentes no *corpus* original e nas entrevistas transcritas automaticamente.

Palavras-chave: Universal Dependencies; Sintaxe; Linguística de corpus; reconhecimento automático de fala.

1 Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Sertãozinho, São Paulo, Brasil; claudias84@gmail.com; https://orcid.org/0009-0003-9388-4297

2 Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo, Brasil; otovale@ufscar.br; https://orcid.org/0000-0002-0091-8079

3 Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo, Brasil; gabiwick@gmail.com; https://orcid.org/0000-0002-7332-4482

Automatic transcription of interviews and Universal Dependencies annotation in the Roda Viva Corpus

Abstract

This article presents research on the automatic transcription of four interviews extracted from the Roda Viva Corpus, which consists of 713 interviews from the Roda Viva Program, on TV Cultura. The original interviews were transcribed by journalists, thus acquiring the status of written text, and also presenting interventions, such as encyclopedic information about facts and people mentioned. In order to work with oral text, this research carried out a pilot work of automatic transcription of four of these interviews, using the Whisper tool. Subsequently, the interviews were automatically annotated with the formalization of Universal Dependencies and manually reviewed by the Arborator Grew ElizIA tool. Through this work, it was possible to note the syntactic differences present in the original corpus and in the automatically transcribed interviews.

Keywords: Universal Dependencies; Syntax; Corpus linguistics; automatic speech recognition.

Introdução

As pesquisas com *corpus* são utilizadas há algum tempo na área de Processamento de Línguas Naturais (PLN) e são fundamentais para que fenômenos linguísticos possam ser observados em seu contexto de uso.

Um tipo especial de *corpus* são aqueles formados por língua oral, como entrevistas, por exemplo, nos quais podem ser observadas ocorrências da língua em uso, com maior ou menor nível de formalidade.

O trabalho tem como objetivo estudar a transcrição automática de quatro dentre as 713 entrevistas que formam o Corpus Roda Viva (Miranda Jr. *et al.,* 2024). De fato, as transcrições que se encontram no Corpus Roda Viva são transcrições manuais de entrevistas realizadas no Programa Roda Viva, da TV Cultura, de 1986 a 2009, feitas por jornalistas. Essas entrevistas apresentam informações adicionais, inseridas pelos transcritores, como informações enciclopédicas sobre fatos e pessoas relatadas e correções de desvios gramaticais, por exemplo. Dessa forma, pode-se afirmar que as transcrições não apresentam exatamente as características de um texto oral, mas sim de um texto jornalístico transcrito. A escolha das quatro entrevistas não se deu ao acaso. Foram escolhidas de acordo com um critério de diversidade que fosse de uma fala mais formal, como a entrevista de Mauricio de Souza, passando por uma entrevista política, que também tinha um alto grau de formalidade, que é o caso da entrevista de Benedita

da Silva (então governadora do Rio de Janeiro), chegando a uma entrevista com menos formalidade no caso do jogador de futebol Edmundo, até uma fala bem característica da periferia paulistana na entrevista de Mano Brown.

A fim de se trabalhar com um texto realmente oral, optou-se, neste trabalho, por realizar a transcrição automática das entrevistas por meio de um ASR chamado Whisper (Radford *et al.,* 2023). Essa ferramenta apresentou alguns problemas na transcrição, que serão apresentados com mais detalhe neste artigo.

Após a transcrição automática das quatro entrevistas e a correção manual dos problemas apresentados, realizou-se a anotação automática delas com as etiquetas da Universal Dependencies (De Marneffe *et al.*, 2021) e uma posterior revisão manual. As entrevistas anotadas e revisadas serão inseridas no *corpus* Porttinari (Pardo *et al*., 2021), a fim de compor a parte oral desse grande *corpus*. Essa escolha se dá em função do formalismo bastante transparente que oferece o quadro das Universal Dependencies.

Na próxima seção será apresentada a fundamentação teórico-metodológica utilizada neste trabalho, com ênfase no gênero oral entrevista e suas características, descrição do Corpus Roda Viva e Corpus Roda Viva TW, ferramentas utilizadas, como o Whisper, parser PortParser, Arborator Grew ElizIA e o formalismo utilizado nas anotações: a Universal Dependencies.

Na seção 3 são apresentados os resultados do trabalho, com ênfase na comparação das transcrições manual e automática das entrevistas.

Por fim, são apresentados os agradecimentos e as referências bibliográficas.

Fundamentação Teórico-Metodológica

Nesta seção serão abordados os fundamentos teórico-metodológicos utilizados neste trabalho.

Gênero oral: entrevista e suas características

Como ponto de partida para se abordar o tema das entrevistas, é necessário definir-se a noção de gênero discursivo. Este trabalho parte da perspectiva de Bakhtin (2016), para quem os gêneros discursivos se constituem a partir de três elementos: o conteúdo temático, o estilo e a construção composicional, além da atividade humana e das situações de interação da vida social dos indivíduos.

Os gêneros discursivos podem ser escritos ou orais, como as entrevistas, sobre as quais Marcuschi (2005, p. 15) destaca alguns pontos importantes, como: a) interação entre pelo menos dois falantes; b) ocorrência de pelo menos uma troca de falantes; c) presença de uma sequência de ações coordenadas; d) execução numa identidade temporal.

Relacionada à oralidade, destaca-se a conversação, através da qual os gêneros orais existem e ocorrem nas diversas situações comunicativas, como as entrevistas, os debates, as notícias. Segundo Marcuschi (1988, p. 319-320), conversação pode ser definida "como uma interação centrada, da qual participam pelo menos dois interlocutores que se revezam, tomando cada qual pelo menos uma vez a palavra, dando-se o evento comunicativo em uma identidade temporal".

Um conceito importante relacionado à conversação é o **tópico discursivo**, que pode ser definido como o ponto de partida de uma conversa. Segundo Lira (2020), para desenvolver o tópico discursivo, o falante considera seus interlocutores, ou seja, suas representações, reações, intenções, expectativas, grau de intimidade, grau de concordância ou discordância sobre o que está sendo discutido no discurso.

Outro conceito importante é o **turno conversacional**, que tem como função organizar o texto oral, para que haja uma progressão lógica das ideias. Em relação aos turnos, ressalta-se que a conversação pode ser de dois tipos: assimétrica e simétrica. Melo Júnior (2016, p. 15) define que, na conversação assimétrica, "há uma hierarquia linguisticamente marcada entre os interactantes de uma situação discursiva, em que um dos participantes do evento de fala detém o poder da palavra e comanda o turno conversacional". Já "as relações simétricas acontecem quando dois interactantes têm o mesmo poder de interagir ou o mesmo poder da palavra" (Melo Júnior, 2016, p. 17).

Ainda em relação à conversação, outro conceito essencial são os **pares adjacentes,** definidos por Marcuschi (2007, p. 35) como "uma sequência de dois turnos que co-ocorrem e servem para a organização local da conversação". Segundo Lira (2020), alguns exemplos dessas sequências são definidos como saudação/saudação, convite/recusa; agradecimento/aceitação; pergunta/resposta.

Finalmente, um último conceito relacionado à conversação que pode ser citado são os **marcadores conversacionais.** Segundo Urbano (1997, p. 81), eles são elementos de "variada natureza, estrutura, dimensão, complexidade semântico-sintática, aparentemente supérfluos ou até complicadores, mas de indiscutível significação e importância para qualquer análise de textos orais e para sua boa e cabal compreensão".

Marcuschi (1986) apresenta a seguinte classificação em relação aos marcadores conversacionais:

1. **MC simples**: é o marcador que se realiza com um só lexema ou um paralexema, como as interjeições, os advérbios, os verbos, os adjetivos, as conjunções, os pronomes, entre outros.

2. **MC composto:** de caráter sintagmático, com grande tendência à estereotipia e com pouca alteração morfológica no tipo produzido.

3. **MC oracional:** trata-se de pequenas orações, podendo se apresentar em todos os tempos e formas verbais ou modos oracionais (assertivo, indagativo, exclamativo).

4. **MC prosódico:** é o MC formado com recursos prosódicos e normalmente utilizado com algum MC verbal. Encontram-se, nesse contexto, a entonação, a hesitação, o tom de voz, entre outros (Marcuschi, 1986, p. 290-291).

Segundo Lira (2020), o texto oral apresenta alguns mecanismos de textualidade característicos, como:

a) **Hesitação** – segundo Marcuschi (2006), a hesitação se constitui nos aspectos formais, cognitivos e interacionais. Ela é caracterizada por repetição de palavras, como as pausas e os alongamentos (de vogais, consoantes ou sílabas). O falante se utiliza da hesitação para reformular sua interação, a fim de se comunicar de forma mais eficiente;

b) **Repetição** – é utilizada para a estruturação do discurso oral e também demonstra características de um planejamento presencial, com características de um texto espontâneo. Segundo Koch (2005, p. 145):

> A repetição é particularmente constitutiva do discurso conversacional, no qual os parceiros, conjuntamente e passo a passo, constroem o texto, elaboram as ideias, criam, preservam e negociam as identidades, de tal forma que o texto, de maneira icônica, vai refletir essa atividade de co-produção.

Koch (2017, p. 84) enfatiza que "[...] a repetição de itens lexicais tem por efeito trazer ao enunciado um acréscimo de sentido que ele não teria se o item fosse usado somente uma vez".

c) **Paráfrase** – Segundo Lira (2020), a paráfrase pode ser definida como a retomada explícita e consciente de outros textos, ou seja, o texto parafraseando não se diferencia do outro, mas sim compartilha semelhanças. As paráfrases podem ser introduzidas por "ou seja", ou "isto é", algumas vezes.

d) **Correção** – Pode ser definida como uma formulação retrospectiva, que exprime contraste. Fávero *et al.* (2006, p. 258) definem a correção como "um enunciado linguístico que reformula um anterior, considerado 'errado' aos olhos de um dos interlocutores".

As entrevistas fazem parte do gênero oral e são caracterizadas por um monitoramento maior dos falantes. Segundo Costa (2023), esse gênero oral pode apresentar vários subtipos, como entrevista televisiva, radiofônica, de emprego, de seleção, pingue-pongue, entre outras.

Conforme Costa (2023), na construção do gênero entrevista, é necessário levar em consideração razões que vão desde sua situação de produção, como o contexto, o propósito comunicativo, o nível de formalidade e o público-alvo, como também sua recepção.

De acordo com Medina (1990), a entrevista é uma forma de interação social entre duas ou mais pessoas, que tem como alvo obter e difundir informações através dos seus participantes, que são o entrevistador e o entrevistado. Ele considera esse gênero como uma técnica de obtenção de informação e de interação entre os participantes.

De acordo com Minayo (2010, p. 280):

> A entrevista é considerada uma modalidade de interação entre duas ou mais pessoas. Essa pode ser definida como a técnica em que o investigador se apresenta frente ao investigado e por meio de perguntas formuladas busca a obtenção dos dados que lhe interessa. É uma conversa a dois, ou entre vários interlocutores, realizada por iniciativa do entrevistador, destinada a construir informações pertinentes para o objeto de pesquisa, e abordagem pelo entrevistador, de temas igualmente pertinentes tendo em vista este objetivo.

Segundo Costa (2023), a entrevista deve ser bem planejada, visando alcançar o seu objetivo principal, que é a obtenção de informações e, por isso, as perguntas devem ser bem formuladas, adequadas à situação e ao tema abordado, levando em consideração o pensamento do entrevistado.

Lira (2020, p. 25) aponta que a entrevista televisiva tem como particularidade o fato de possibilitar a visualização de imagens, o que gera uma maior aproximação entre o entrevistador e o entrevistado e os telespectadores e isso colabora com a interação social entre os sujeitos.

Costa (2023), em seu trabalho, apresenta uma tabela com algumas características das entrevistas, que são apresentadas no Quadro 1:

Quadro 1. Características das entrevistas

CARACTERÍSTICAS DAS ENTREVISTAS	
GÊNERO ENTREVISTA	Número de participantes: deve haver a presença de no mínimo um entrevistador e um entrevistado, porém, em algumas entrevistas, pode conter mais de uma pessoa para cada uma dessas categorias.
	Forma da interação entre entrevistador e entrevistado: a relação entre os interlocutores, ou seja, a condução, a troca de turnos e as interrupções.
	Relação do entrevistado com o tema abordado na entrevista: pode ser sobre sua vida ou um tema de sua área de atuação.
	Enquadramento da fala ao nível de formalidade: entrevistador e entrevistado devem adequar suas falas à situação comunicativa e ao nível de formalidade que ela exige.
	Elementos paralinguísticos e cinésicos: são aspectos como o tom da voz, o ritmo e as pausas na fala. E movimentos corporais, expressões faciais e gestos.

Fonte: Costa (2023, p. 13)

Após serem elencadas as principais características sobre as entrevistas, na subseção que segue, será apresentado o Corpus Roda Viva, composto por entrevistas televisivas.

Corpus Roda Viva

O Corpus Roda Viva[4] (Miranda *et al*., 2024) foi criado no âmbito do Projeto Memória Roda Viva, em uma parceria da Fundação Padre Anchieta, Fapesp e Unicamp e apresenta 713 entrevistas do Programa Roda Viva, da TV Cultura, de janeiro de 1986 até julho de 2009. Ele apresenta verbetes, referências, fotos e vídeos. Segundo Miranda *et al*. (2024), esse portal foi alvo de apenas dois trabalhos, Botin (2016) e Pacheco (2020), que tiveram como foco análises linguísticas teóricas. De acordo com Miranda *et al*. (2024), do total de 713 entrevistas do Corpus Roda Viva, apenas 364 delas possuem sua versão em vídeo, sendo que 308 delas (446h 18min 49s) são em português do Brasil e o restante também em outras línguas como inglês, espanhol, português europeu, francês e italiano. As entrevistas são divididas em cinco grandes temas: Ciências, Cultura, Esporte, Economia e Política.

4 Acesso por: rodaviva.fapesp.br

Atualmente, o Corpus Roda Viva possui duas versões: a V0.1 apresenta um total de 517.256 sentenças, 9.859.582 *tokens* e 2.633.400 *types*. As entrevistas foram transcritas por jornalistas que fizeram algumas modificações, tornando o *corpus* mais formal e com características de um texto escrito. As inserções textuais no *corpus* são colocadas em colchetes. Alguns exemplos de inserções são:

A) complementação de palavras omitidas durante a fala: *""Eu acho que* ***[ele]*** é o melhor do mundo como chargista»;

B) direcionamentos de como o fluxo conversacional se desenrola: "*Se voce* ..., na *eleição..., poderia fazer...* ***[falando junto com o Markun e concordando com ele]";***

C) explicação de siglas: *PIB* ***[Produto Interno Bruto];***

D) explicação enciclopédica sobre um fato ou alguém: "*O Paulinho* ***[Paulinho da Viola, cantor e compositor]*** *gravou o quê?"*

Na segunda versão do Corpus, a V0.2, foram removidas as intervenções acrescentadas pelos jornalistas no texto original. Essa exclusão representou uma redução de 2,5% de sentenças e 4,6% de *tokens*, segundo apresentam Miranda *et al.* (2024). Ambas as versões estão disponíveis em dois formatos: uma compilação de arquivos CSV (com uma entrevista por arquivo) e um arquivo JSON com todas as entrevistas. O arquivo CSV contém 5 colunas com as informações da data, nome da entrevista, ordem de fala, nome do falante e o texto.[5]

Alguns objetivos da construção desse *corpus* são: ter-se um registro importante da história recente, a preservação das entrevistas, acesso livre a todo conteúdo e retroalimentação das pautas do programa.

Corpus Roda Viva TW

O Corpus Roda Viva TW é um projeto piloto composto por quatro entrevistas extraídas do Corpus Roda Viva (Miranda *et al.*, 2024), totalizando 4.024 sentenças e 69.377 tokens. As quatro entrevistas foram escolhidas no intuito de apresentarem uma possível diversidade sintática, pois os entrevistados são pessoas com formações bem diversas, sendo um jogador de futebol, um *rapper*, um escritor de revista em quadrinhos e uma governadora de estado.

Como já foi salientado na subseção anterior, as entrevistas do Corpus Roda Viva foram transcritas manualmente por jornalistas, apresentando, por isso, informações adicionais. Com o objetivo de ser mais fiel à fala, decidiu-se transcrever automaticamente quatro

5 Arquivos disponíveis no GitHub: https://github.com/<ANONYMIZED>/Roda-Viva

entrevistas (em vistas do curto período de tempo disponível para o trabalho), por meio de um Sistema de Reconhecimento Automático de Fala (ASR) chamado Whisper (Radford *et al*., 2023), que será apresentado com mais detalhes na próxima subseção.

Após isso, o Corpus Roda Viva TW foi anotado, também automaticamente, com as etiquetas da Universal Dependencies (De Marneffe *et al*., 2021), por meio do parser PortParser (Lopes *et al*., 2024) e revisado manualmente com relação à anotação, com o uso da ferramenta Arborator Grew-ElizIA (Guibon *et al*., 2020).

Nas subseções seguintes, serão apresentadas as ferramentas utilizadas no processamento do Corpus Roda Viva TW, bem como o formalismo da Universal Dependencies.

ASR Whisper

Os Sistemas de Reconhecimento Automático de Fala têm sido muito utilizados nos dias de hoje para realizarem inúmeras tarefas. Com vistas a transcrever de forma automática as quatro entrevistas que formam o Corpus Roda Viva TW, foi utilizado neste trabalho o ASR Whisper (Radford *et al*., 2023).

De acordo com Radford *et al*. (2023), o Whisper se concentra em ampliar o escopo do pré-treinamento fracamente supervisionado, além do reconhecimento de fala apenas em inglês, sendo tanto multilíngue quanto multitarefa. Das 680.000 horas de áudio do treinamento da ferramenta, 117.000 horas abrangem 96 outras línguas. O conjunto de dados também inclui 125.000 horas de dados de tradução X en.

A utilização de ASR traz algumas considerações éticas, uma vez que esses sistemas costumam gravar e armazenar áudios para treinar modelos e aprimorar-se, o que pode trazer preocupações com relação à invasão de privacidade, retenção excessiva e usos secundários de dado. Para sanar essa questão, é necessário adotar-se consentimento informado, oferecendo clareza sobre como e por que os dados de voz são coletados, por exemplo.

Com relação à variação linguística, observa-se que a cobertura limitada de diferentes sotaques e dialetos no treinamento dos ASR pode frustrar e excluir grupos menos representados. Para que isso não ocorra, os sistemas precisam ser treinados com conjuntos de dados diversificados.

O Whisper sugere que a simples ampliação do pré-treinamento fracamente supervisionado tem sido subestimada até agora para reconhecimento de fala. Os autores alcançaram esses resultados sem a necessidade de técnicas de auto-supervisão ou auto-treinamento, que têm sido uma constante no trabalho recente em reconhecimento

de fala em larga escala, e demonstram como o simples treinamento em um conjunto de dados supervisionado grande e diversificado e o foco na transferência zero-shot podem melhorar significativamente a robustez de um sistema de reconhecimento de fala.

Os vídeos das entrevistas extraídos do YouTube foram submetidos ao Whisper, e a ferramenta gerou um arquivo .txt com o texto transcrito. Esses textos apresentaram alguns problemas, como segmentação errada de frases e alguns erros de transcrição de palavras, principalmente com estrangeirismos e por má dicção dos falantes. Para corrigir isso, todos os textos foram revisados manualmente antes de anotá-los com o parser PortParser (Lopes *et al.*, 2024).

No Quadro 2 são apresentados os números relativos aos erros de identificação de palavras observados em cada uma das entrevistas, bem como um exemplo de erro e seu contexto de ocorrência.

Quadro 2. Números dos erros de identificação de palavras e entrevistas-fonte

Número de erros	Entrevista fonte	Exemplo	Contexto
21	Mano Brown	Afinar da (a finada)	*A finada classe média*
31	Edmundo	Zazinho (Zezinho)	*O seu Zezinho Mansur*
14	Maurício de Souza	Passeato (passeata)	*a criançada fez passeata em Brasília*
14	Benedita de Silva	Disseu (Dirceu)	*O Lula, pessoalmente, o Zé Disseu, convenceram a senhora a... a assumir e agora já estaria certa a sua participação no governo?*

Fonte: Elaboração própria

Parser PortParser

Após a transcrição automática do Corpus Roda Viva TW e a correção manual dos problemas apresentados pelo ASR Whisper, iniciou-se a etapa de anotação automática do *corpus* com as etiquetas da Universal Dependencies. A ferramenta escolhida para essa tarefa foi o parser PortParser (Lopes *et al.*, 2024).

De acordo com Lopes *et al.* (2024), o Portparser supera os sistemas atuais para textos jornalísticos em português brasileiro. Seguindo o *framework* de Dependências Universais (UD), o modelo foi construído utilizando um *corpus* manualmente anotado recentemente

lançado (Porttinari-base) para treinamento. Os autores testaram diferentes métodos de análise sintática e exploraram configurações de parâmetros com o objetivo de propor um modelo altamente preciso, abrangendo não apenas a anotação de dependências, mas também a marcação de Part-of-Speech, a identificação de lemmas e as características morfológicas relacionadas. O melhor modelo alcançou cerca de 99% de precisão na marcação de Part-of-Speech, lemmas e características morfológicas, com cerca de 95% de precisão na anotação de dependências, superando sistemas conhecidos para o português em até 7% de precisão. Os autores também realizaram uma análise de erros do modelo proposto para mostrar as limitações atuais e os desafios para trabalhos futuros.

O *parser* se mostrou muito capacitado para anotar o *corpus* trabalhado, cometendo apenas alguns erros de anotações com relação a estruturas de fala, como vocativos, marcadores discursivos, objetos diretos antepostos ao verbo, hesitações e truncamentos.

Não foi possível saber quantas correções foram realizadas em cada sentença, pois teria sido necessário não salvar nenhuma alteração nas sentenças em que não foram identificados problemas, o que não foi feito, para que fosse possível se calcular quantas sentenças sofreram alterações e quantas não. Somado a isso, por limitações da ferramenta, também não foi possível se contabilizar quantos *tokens* sofreram uma revisão com relação à anotação.

Arborator Grew-ElizIA

A fim de se revisar a anotação automática do *corpus* realizada pelo *parser* PortParser (Lopes *et al.,* 2024), foi utilizada a ferramenta Arborator-Grew ElizIA (Guibon *et al.*, 2020), que permite visualizar as sentenças anotadas e fazer as correções necessárias.

De acordo com Guibon *et al.* (2020), o Arborator-Grew combina as funcionalidades de duas ferramentas pré-existentes: Arborator e Grew. Arborator é uma ferramenta colaborativa amplamente utilizada para anotação gráfica *online* de árvores de dependência. Grew é uma ferramenta para consulta e reescrita de grafos, especializada em estruturas necessárias em PLN, ou seja, árvores e grafos de dependência sintática e semântica. O Arborator-Grew é um redesenho completo e uma modernização do Arborator, substituindo seu próprio armazenamento interno de banco de dados por uma nova API do Grew, que adiciona uma poderosa ferramenta de consulta às funcionalidades existentes de criação e correção de banco de dados de árvores do Arborator. Isso inclui controle de acesso complexo para anotação paralela por especialistas e crowdsourcing, visualização de comparação de árvores e vários modos de exercício para ensino e treinamento de anotadores. O Arborator-Grew abre novos caminhos para a criação, atualização, manutenção e curadoria coletiva de bancos de dados de árvores sintáticas e bancos de grafos semânticos.

A fim de se realizar a revisão da anotação no Arborator-Grew ElizIA, dois anotadores foram selecionados, sendo um com vasta experiência sobre a anotação Universal Dependencies e outro iniciante, o qual passou por um treinamento prévio antes da tarefa. Os dois anotadores se reuniam regularmente a fim de debaterem sobre casos complexos na anotação. Dessa forma todas as sentenças foram sendo revisadas ao longo do trabalho.

Como forma de se calcular a concordância dos anotadores, foi feito um recorte do *corpus* formado por 200 sentenças aleatórias. O resultado desse cálculo pode ser observado na Tabela 1, em que se nota a porcentagem da concordância entre os anotadores com relação às formas, lemas, etiquetas morfossintáticas, características, núcleos e relações de dependência. Nesses dados, percebe-se que a concordância foi baixa em relação à atribuição dos núcleos e das relações de dependência.

Tabela 1. Porcentagem de concordância entre os anotadores

Tipo de Token	Porcentagem de concordância
FORM (forma)	98,81%
LEMMA (lema)	97,92%
UPOS (etiqueta morfossintática)	95,56%
FEAT (características)	95,24%
HEAD (núcleo)	47,02%
DEPREL (relação de dependência)	55,65%

Fonte: Elaboração própria

Universal Dependencies

Como já foi citado, as entrevistas do Corpus Roda Viva TW foram anotadas com as etiquetas do modelo da Universal Dependencies (UD) (De Marneffe *et al*., (2021), o qual almeja realizar uma anotação gramatical consistente (etiquetas morfossintáticas, características morfológicas e dependência sintática), entre línguas humanas diferentes. Esse modelo é um esforço colaborativo de cerca de 500 pessoas que produziram quase 200 *treebanks* para aproximadamente 100 línguas.

Neste momento, a UD possui dezessete etiquetas morfossintáticas ou Part-of-Speech (PoS) *tags*, como: NOUN: substantivo, DET: determinante e PRON: pronome.

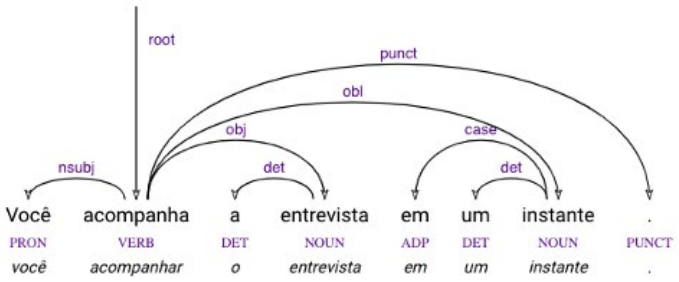
Além disso, a UD também possui 37 etiquetas de relações de dependência – *deprel* (de dependency relation), como PUNCT – pontuação; NSUBJ – sujeito; NMOD – modificador nominal. Uma *deprel* é uma relação que liga dois a dois os elementos (*tokens*) de uma

sentença. Um deles é chamado de *head* (núcleo), que é sempre uma palavra de conteúdo (verbo, substantivo, adjetivo, pronome, numeral e advérbio) e o outro é chamado de dependente.

Normalmente o predicado da oração principal é a raiz de uma sentença, marcada como dependente da *deprel root*. Os arcos das relações não devem se cruzar, pois a atribuição de relações de dependência deve observar o princípio da projetividade.

A Figura 1 apresenta um exemplo de uma sentença anotada com relações de dependência UD.

Figura 1. Exemplo de árvore de dependências anotada com etiquetas da UD



Fonte: Corpus Roda Viva TW

Após a apresentação do arcabouço teórico-metodológico utilizado nesta pesquisa, passa-se, na seção 3 a apresentar-se os resultados do trabalho.

Análise e resultados

Nesta seção serão apresentados alguns resultados de comparação entre trechos de entrevistas do Corpus Roda Viva e do Corpus Roda Viva TW, através dos quais é possível perceber que o primeiro *corpus* não apresenta características de um *corpus* de língua falada, devido à sua transcrição manual. Já o Corpus de Roda Viva TW, transcrito manualmente, mantém essas características.

O Quadro 3 apresenta dois trechos de uma das entrevistas, sendo um deles transcrito manualmente (Corpus Roda Viva) e o outro automaticamente (Corpus Roda Viva TW). Nesses trechos é possível perceber que algumas marcas de oralidade são perdidas na transcrição manual, como a repetição de palavras e a presença de marcadores discursivos, bem como são feitas algumas correções gramaticais, como a inserção de pronomes omitidos na fala.

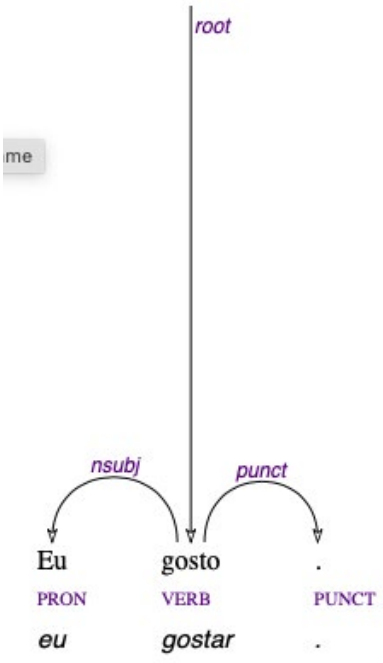
Quadro 3. Diferenças entre a transcrição manual e automática

Transcrição Manual	Transcrição Automática
Edmundo: Eu gosto. Número de camisa, eu nunca tive nenhuma preferência. Coincidentemente, eu marquei pela camisa sete. Fui para o Flamengo, gostaria também de ter mudado quando eu cheguei no Flamengo e lá foi feito uma votação e ganhou por 92% camisa sete. E aqui, o Marcelinho é o dono da camisa sete e gosta de jogar com ela, eu acho que não tinha nem como eu pensar em jogar com a camisa sete. E aí foram escolhidos outros dois números, caiu o número oito. Eu me simpatizo bem com o número oito.	Eu gosto, **gosto.** Número de camisa, nunca tive nenhuma preferência, **né?** Coincidentemente, marquei pela camisa 7. Fui para o Flamengo, gostaria também de ter mudado quando cheguei no Flamengo. E lá foi feita uma votação e ganhou por 92% a camisa 7. E aqui o Marcelinho é o dono da camisa 7 e gosta de jogar com **a camisa 7.** Eu acho que não tinha nem como eu pensar em jogar com a camisa 7. E aí foram escolhidos os outros dois números, caiu o número 8, e eu me simpatizo bem com o número 8.

Fonte: Elaboração própria

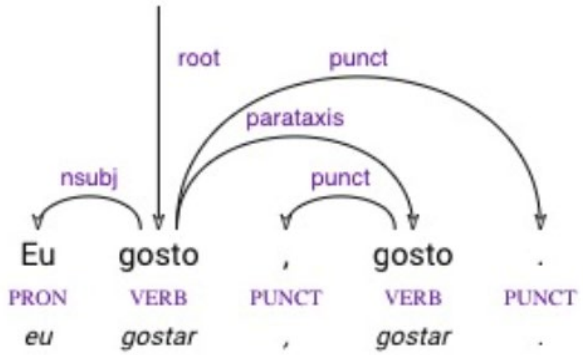
As Figuras 2 e 3 apresentam as diferenças de anotação com as etiquetas da UD de duas sentenças presentes nos trechos apresentados na Tabela 2. Verifica-se que também a anotação de dependências se altera nos dois tipos de transcrição.

Figura 2. Anotação UD de uma sentença transcrita manualmente



Fonte: Corpus Roda Viva

Figura 3. Anotação UD de uma sentença transcrita automaticamente



Fonte: Corpus Roda Viva TW

Por meio desses exemplos, é possível perceber as diferenças que existem entre o *corpus* com as entrevistas transcritas manualmente (Corpus Roda Viva) e o *corpus* com as entrevistas transcritas automaticamente (Corpus Roda Viva TW), pois este último apresenta marcas da oralidade, enquanto o anterior já não as apresenta, motivo pelo qual optou-se pela transcrição automática das entrevistas neste trabalho.

O objetivo final deste trabalho é anotar automaticamente as outras entrevistas do corpus Roda Viva, a fim de se inserir uma porção de *corpus* oral no projeto Porttinari (Pardo *et al.*, 2021), que já possui outros gêneros de textos escritos como *tweets* do mercado financeiro, *reviews* de *e-commerce*, entre outros.

Referências

BAKHTIN, M. *Os gêneros do discurso.* Paulo Bezerra (Organização, Tradução, Posfácio e Notas); Notas da edição russa: Seguei Botcharov. São Paulo: Editora 34, 2016. 164p.

BOTIN, L. M. *Ciência e tecnologia em debate:* uma análise das entrevistas do programa Roda Viva, da TV Cultura. 2016. Tese (Doutorado em Letras), Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, Brasil, 2016.

COSTA, M. B. Uma análise do gênero entrevista: identificação de seus tipos, características e especificidades. *Portal UFERSA,* 2023.

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal dependencies. *Computational linguistics*, v. 47, n. 2, p. 255-308, 2021.

FAVERO, L. L. *et*. *al*. *Gramática do Português Culto Falado no Brasil*: construção do texto falado. v.1. Campinas: Editora da Unicamp, 2006.

GUIBON, G.; COURTIN, M.; GERDES, K.; GUILLAUME, B. When collaborative treebank curation meets graph grammars: arborator with a grew back-end. *Proceedings of the 12th Language Resources and Evaluation Conference,* Marseille, France, European Language Resources Association, p. 5293- 5302, maio 2020.

KOCH, I. V. *Introdução* à *linguística textual.* 2. ed. São Paulo: Contexto, 2017.

KOCH, I. V. *et. al*. *Referenciação e discurso*. São Paulo: Contexto, 2005.

LIRA, L. S. *Análise de aspectos orais no gênero entrevista televisiva*. 2020. TCC (Graduação) – Curso de Letras – Língua Portuguesa, Universidade Federal de Alagoas, Arapiraca, 2020.

LOPES, L.; PARDO, T.A.S. Towards Portparser – a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR),* p. 401-410. May. 13-15. Disponível em: https://aclanthology.org/2024.propor-1.41. Acesso em: 04 nov. 2024.

MARCUSCHI, L. A*. Análise da Conversação.* 6. ed. São Paulo: Ática, 2007.

MARCUSCHI, L. A. *et. al.* Fenômenos Intrínsecos da Oralidade: a hesitação. *In:* MARCUSCHI, L. A. *et. al. Gramática do português falado:* construção do texto falado. Campinas: Editora da Unicamp, 2006.

MARCUSCHI, L. A. *et. al.* Oralidade e ensino de língua: uma questão pouco "falada". *In:* MARCUSCHI, L. A. *et. al. O livro didático de português:* múltiplos olhares*.* 3. ed. Rio de Janeiro: Lucerna, 2005.

MARCUSCHI, L. A. *Questões atuais na Análise da Conversação.* Recife: ANPOLL, 1988.

MARCUSCHI, L. A. *Análise da conversação*. São Paulo: Ática, 1986.

MEDINA, C. A. *Entrevista*: o diálogo possível. 3. ed. São Paulo: Ática, 1990.

MELO JÚNIOR, J. N. B. *Aspectos textuais e conversacionais na entrevista oral no radiojornalismo alagoano.* 2016. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Programa de pós-graduação profissional em Letras e Linguística, Universidade Federal de Alagoas, Maceió, 2016.

MINAYO, M. C. S. Técnicas de pesquisa: entrevista como técnica privilegiada de comunicação. *In*: MINAYO, M. C. S. *O desafio do conhecimento*: pesquisa qualitativa em saúde. 12. ed. São Paulo: Hucitec, 2010. p. 261- 297

MIRANDA Jr., I.; PEDRO, G. W.; BARROS, C. D.; VALE, O. A. Roda Viva Boundaries: an overview of an audio-transcription corpus. *Proceedings of the 16th International Conference on Computational Processing of Portuguese,* vol, 2, p. 165-169, mar. 2024.

PACHECO, P. H. *A construção "acontece que" no português brasileiro contemporâneo:* uma análise baseada no uso. 2020. Dissertação (Mestrado em Letras) – Universidade Federal Fluminense, Niterói, Brasil, 2020.

PARDO, T. A. S.; DURAN, M. S.; LOPES, L.; DI FELIPPO, A.; ROMAN, N. T.; NUNES, M. G. V. Porttinari – a large multi-genre treebank for Brazilian Portuguese. *In*: *Proceedings of the XIII Symposium in Information and Human Language (STIL),* p. 1-10, 2021.

RADFORD, A; KIM, J. W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning,* PMLR 202, p. 28492-28518, 2023.

URBANO, H. Marcadores conversacionais. *In*: PRETI, D. (org.). *Análise de textos orais.* 3. ed. São Paulo: Humanitas Publicações – FFLCH/USP, 1997. p. 81-101.