UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE ESTUDOS DA LINGUAGEM

LUANA MORO

TREINAMENTO LINGUÍSTICO DE *SOFTWARE* NA PÓS-EDIÇÃO DE TRANSCRIÇÃO E TRADUÇÃO AUTOMÁTICA EM CURSOS DE EDUCAÇÃO A DISTÂNCIA

CAMPINAS

2019

LUANA MORO

TREINAMENTO LINGUÍSTICO DE *SOFTWARE* NA PÓS-EDIÇÃO DE TRANSCRIÇÃO E TRADUÇÃO AUTOMÁTICA EM CURSOS DE EDUCAÇÃO A DISTÂNCIA

Dissertação apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas para a obtenção do título de Mestra em Linguística Aplicada, na área de Linguagem e Sociedade.

Orientador: Prof. Dr. Rodrigo Esteves de Lima Lopes
Co orientador: Prof. Dr. Daniel Yokoyama Sonoda

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA LUANA MORO, E ORIENTADA PELO PROF. DR. RODRIGO ESTEVES DE LIMA LOPES.

CAMPINAS

2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Estudos da Linguagem
Leandro dos Santos Nascimento - CRB 8/8343

Moro, Luana, 1991-
M828t Treinamento linguístico de software na pós-edição de transcrição e tradução automática em cursos de educação a distância / Luana Moro. – Campinas, SP : [s.n.], 2019.

Orientador: Rodrigo Esteves de Lima-Lopes.
Coorientador: Daniel Yokoyama Sonoda.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Estudos da Linguagem.

1. Linguística de corpus. 2. Processamento eletrônico de dados. 3. Inteligência artificial. 4. Educação a distância. I. Lopes, Rodrigo Esteves de Lima. II. Sonoda, Daniel Y. III. Universidade Estadual de Campinas. Instituto de Estudos da Linguagem. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Linguistic training for software in post-edition of automatic transcription and translation machines in distance learning
Palavras-chave em inglês:
Corpora (Linguistics)
Electronic data processing
Artificial intelligence
Distance education
Área de concentração: Linguagem e Sociedade
Titulação: Mestra em Linguística Aplicada
Banca examinadora:
Rodrigo Esteves de Lima Lopes
Roberto Carlos Assis
Marcia Veirano Pinto
Data de defesa: 30-08-2019
Programa de Pós-Graduação: Linguística Aplicada

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-0745-0552
- Currículo Lattes do autor: http://lattes.cnpq.br/4791664021671125

BANCA EXAMINADORA:

Rodrigo Esteves de Lima Lopes

Roberto Carlos Assis

Marcia Veirano Pinto

IEL/UNICAMP
2019

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação do IEL.

"Mas é preciso ter força
É preciso ter raça
É preciso ter gana sempre"

Milton Nascimento – Maria, Maria

Dedico este trabalho à todas as mulheres
fortes que se empoderam diante da ciência,
da vida e dos preconceitos.

Para a mulher mais forte
que já tive o prazer de conhecer, Tereza.

AGRADECIMENTOS

Aos meus guias e a Deus, que sempre seguram minhas mãos e me direcionam para o caminho do bem.

À minha família, em especial meus avós que já cuidam de mim do outro plano, Hélvio e Tereza. Aos meus pais, Marcelo e Helis Regina, pela vida, por me mostrarem o que é resiliência e perdão ao longo dos percalços da vida. Ao meu irmão Samuel, que me mostrou o que é sentir o próprio coração fora do corpo. À Fernanda, que nunca saiu do meu lado, me motivou e ouviu meus choros e alegrias (e sempre me forneceu os melhores suprimentos acadêmicos da turma). À minha tia Daniela, por sempre me apoiar e por ter me dado aquele dicionário que mudou tudo. Ao meu tio Felippe, por me inspirar em sua genialidade.

Ao Lucas, por ser minha família, por despertar o melhor de mim, por me apoiar em todas as aventuras dessa vida (e das outras também) e por me amparar quando eu mesma já não era o suficiente.

À Fernanda Bacellar, por ter me mostrado quem eu realmente sou, pelo amparo, orientação na vida e carinho.

Ao Pecege pelo apoio financeiro, pelo acolhimento e por tudo. Em especial ao Daniel Sonoda, por sempre acreditar em mim e no meu sonho.

Ao meu orientador Rodrigo, que teve uma paciência do tamanho do universo, que me acolheu e orientou com tanto zelo. Aos membros da banca, Prof. Márcia Veirano e Prof. Roberto Assis pelas incríveis considerações que tornaram o curso do meu trabalho ainda mais grandioso.

Ao professor Marcelo Buzato, por sempre ter apoiado minhas maluquices tecnológicas, ter me guiado brilhantemente durante suas aulas incríveis e por suas considerações imprescindíveis na qualificação.

Às minhas amigas Bruna e Aline, por serem minhas irmãs de coração, que sempre me seguraram quando eu estava prestes a cair. Pela cumplicidade.

Aos meus amigos de luta, de RU e feirinha: Daniel, Tiêgo, Izadora, Ludmilla e Gabriel. Especialmente ao Daniel, que sempre esteve do meu lado e que se tornou um amigo para vida. Ao Tiêgo, pela força e playlists incríveis.

Às minhas amigas Marina Sansão e Juliana Goulart, por todo apoio, amor, vinhos, choros e risadas.

Ao meu amigo Rauster, que sempre esteve ao meu lado e por ser o único que sobrou, desde os primórdios do fim do mundo (o que foi em 2012) e da faculdade.

Aos amigos e membros do grupo de pesquisa Mídites, em especial Rodrigo, Marco, Izadora, Maristella e Terezinha, que sempre estiveram ao meu lado com suas contribuições.

A todos que me disseram palavras de força e motivação no processo de escrita. Este não foi fácil.

À minha equipe maravilhosa, que faz a Skylar acontecer todo dia: Lucas, Gabriel, Marcos, Luan, Anna, Rodolfo e Aline.

Aos funcionários do IEL, que sempre fornecem o melhor atendimento.

A todos do Pecege, que sempre me apoiaram incondicionalmente.

RESUMO

Esta pesquisa tem por objetivo elaborar modelos de treinamento para o software de transcrição e tradução automática, Skylar, tendo como embasamento teórico a linguística de *corpus* (LC) (BERBER SARDINHA, 2004) e o processamento de linguagem natural (PLN) (LIDDY, 2001). A Skylar é uma ferramenta de inteligência artificial (IA) que se estabelece no contexto da educação a distância (EAD) e tem por objetivo realizar legendas para uma videoaula através do reconhecimento de voz. Ferramentas que buscam a automatização com o PLN podem contar erros de entendimento através do reconhecimento de voz na geração de seus conteúdos (OTHERO, 2006), o que causa um problema no entendimento do consumidor dessas legendas. Para tanto, este treinamento se fez necessário pelo uso da *Skylar* na transcrição e tradução de videoaulas, pois estes equívocos podem prejudicar o entendimento do aluno que irá assistir a aula com o conteúdo legendado. Para a execução do processo de metodológico, foi selecionada uma videoaula aula de MBA no tema de Agronegócios para análise da transcrição que, como apresentado anteriormente, é o ponto central de análise desta pesquisa. A metodologia é quanti-qualitativa, sendo dividida em três partes: 1) análise quantitativa – aplicação de dois cálculos estatísticos, WER – *Word Error Rate* (KLAKOW; PETERS, 2002) e LD – *Levensthein distance* (LEVENSTHEIN, 1966), para se obter o índice de erros e a distância (a comparação entre os caracteres gerados automaticamente em relação ao conteúdo ideal) entre as letras certas e erradas, respectivamente; 2) análise qualitativa – os dados obtidos nos cálculos serão analisados com o filtro de melhores e piores segmentos, levantando hipóteses e insumos para o desenvolvimento da interpretação utilizando técnicas de linguística do *corpus*; e 3) apuração dos conteúdos em modelos padronizados para o treinamento do *software*, chegando ao objetivo desta pesquisa. Além dos processos citados, durante a aplicação da análise qualitativa, serão isolados termos mais frequentes para análise de contexto, de forma a fornecer ao *software* as áreas de concentração em que a aula pertence. A partir da análise de 406 segmentos de legenda (31 minutos de vídeo), observou-se que o software obteve um resultado positivo, ou seja, a legenda gerada através do reconhecimento de voz foi mais próxima do que o locutor falou, através da análise quantitativa. No entanto, ao serem aplicadas classificações de erros e na elaboração da análise qualitativa, identificou-se a necessidade de um treinamento que se voltasse para a apuração de contexto.

Palavras-chave: Linguística de *Corpus*; Treinamento de *Software*; Transcrição; EAD

ABSTRACT

This research aims to develop training models for the transcription and machine translation software, Skylar, based on corpus linguistics (CL) (BERBER SARDINHA, 2004) and Natural Language Processing (NLP) (LIDDY, 2001) theories. Skylar is an artificial intelligence (AI) tool that is established in the context of distance education (DE) and aims to make subtitles for a video class through voice recognition. Tools that seek automation with NLP can count misunderstandings through speech recognition in the generation of their content (OTHERO, 2006), which causes a problem in consumer understanding of these subtitles. Therefore, this training was necessary using Skylar in the transcription and translation of video lessons, as these mistakes can undermine the understanding of the student who will attend the class with the subtitled content. For methodological process, an MBA video class was selected about Agribusiness for transcription analysis, which, as previously presented, is the central point of analysis of this research. The methodology is quantitative and qualitative, being divided in three parts: 1) quantitative analysis - application of two statistical calculations, WER - Word Error Rate (KLAKOW; PETERS, 2002) and LD - Levensthein distance (LEVENSTHEIN, 1966), to obtain the error index and the distance (the comparison between automatically generated characters versus ideal content) between the right and wrong letters, respectively; 2) qualitative analysis - the data obtained in the calculations will be analyzed with the filter of best and worst segments, raising hypotheses and inputs for the development of interpretation using techniques of corpus linguistics; and 3) verification of the contents in standardized models for software training, reaching the objective of this research. In addition to the cited processes, during the application of qualitative analysis, more frequent terms for context analysis will be isolated to provide the software with the concentration areas in which the class belongs. From the analysis of 406 subtitle segments (31 minutes of video), it was observed that the software obtained a positive result, that is, the subtitle generated through voice recognition was closer to what the speaker spoke, through the analysis. quantitative. However, when applying error classifications and in the elaboration of the qualitative analysis, it was identified the need for training that focused on the context determination.

Keywords: *Corpus* Linguistics; Software training; Transcription; Distance learning

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

SUMÁRIO

INTRODUÇÃO

Este trabalho tem como objetivo o desenvolvimento de um treinamento linguístico para a ferramenta de transcrição e tradução automática, *Skylar*[1], baseando-se em aspectos metodológicos relacionados aos princípios da linguística de *corpus* (BERBER SARDINHA, 2004; BIBER; REPPEN, 2015; SINCLAIR, 1991) utilizando métricas da linguística computacional (KLAKOW; PETERS, 2002; LEVENSTHEIN, 1966). A ferramenta em questão é aplicada no contexto da educação a distância (EAD) (LITTO; FORMIGA, 2011), através da geração de legendas a partir de videoaulas e proveniente do desenvolvimento de aplicações de processamento de linguagem natural (PLN) (LIDDY, 2001). Este PLN aplicado no *software*, foi desenvolvido no âmbito do reconhecimento de voz, incluindo técnicas de *machine learning* para a sintetização da voz e a geração automática de legendas. Sabe-se que as ferramentas de sintetização da voz, também chamados transcritores, por vezes não executam esta tarefa de forma precisa, ou seja, gerando equívocos no material que processam (OTHERO, 2006). Para tanto, este treinamento se fez necessário pelo uso da *Skylar* na transcrição e tradução de videoaulas, pois estes equívocos[2] podem prejudicar o entendimento do aluno que irá assistir a aula com o conteúdo legendado.

Impulsionado pelo desenvolvimento de ferramentas e utilizando a *internet* como meio, o processo de globalização representou e ainda representa mais do que um processo envolvendo trâmites internacionais interligado entre os países, trocas de cultura e etc. que ocasionam uma mudança maior da percepção do mundo. Essa mudança que teve início no final do século XX, conecta nações, ideias e culturas diferentes interligadas por um fator em comum, a disseminação do conhecimento (IANNI, 2001, p.13).

Entendendo que a criação *internet* e suas ferramentas foram um marco na história da humanidade, estabelecendo uma nova virada antropológica (ABREU, 2009, p. 3), esta tornou-se um local de concentração de informação e comunicação, alterando o espaço-tempo de atividades comuns, permitindo que as pessoas buscassem modelos

---

[1] www.*skylar*.ai

[2] É considerado "equívoco" ou "erro" nesta pesquisa, os trechos que não estão em conformidade com o que foi dito pelo locutor no momento da captação da voz, ou até mesmo problemas gramaticais e de contexto por uma falta de linearidade que ocorre na oralidade, não sendo bem executado quando transpomos para o recurso textual.

diferenciados para e obter o conhecimento, geralmente de forma informatizada (MORAN; MASETTO; BEHRENS, 2000, p. 11). Neste cenário, a modificação dos modelos educacionais mediados por meios digitais se desenvolveu como uma interação diferenciada às práticas tradicionais, na tentativa de atender às necessidades de pessoas que não possuem disponibilidade de tempo para estudar e o encurtamento do tempo das práticas de aprendizado (BUSTAMANTE, 2010, p. 21).

Essa evolução no âmbito da disseminação do conhecimento e na distribuição de conteúdo educacional, fez com que os meios de aprendizado se tornassem mais acessíveis ao longo dos anos. A metodologia de EAD tem tido um impulso relevante nas duas últimas décadas, posicionando esta categoria à frente do desenvolvimento de novas práticas educacionais, de grande interesse público (GARRISON, 2000, p.1). Embora muitos acreditem que a EAD é uma temática nova, ela apresenta indícios de propagação desde 1833, com a comunicação através dos correios (GOMES, 2003, p. 138). Porém, com o desenvolvimento da sociedade globalizada, a transformação digital e o uso da rede de *internet* mediatizando os conteúdos informacionais (GOMES, 2003, p. 138), a prática da EAD foi transformada, sendo regulamentada como ensino regular, e democratizando o acesso à educação por diferentes meios (ALMEIDA, 2002).

Com o surgimento da *internet* como meio de comunicação e as possibilidades latentes na área de educação (GUIMARÃES, 2007, p. 140), a demanda por EAD nos meios digitais realiza um papel importante dentro do processo de globalização. Com o avanço e o aumento dessa demanda nesses meios, a modalidade de EAD passa por diferentes pontos de vista entre as pessoas, que variam entre a inclusão das experiências interculturais, variedade de formas de ensino mediadas por tecnologia, novas iniciativas científicas e ainda, projetos de desenvolvimento internacionais (KNIGHT, 2004).

Com a ideia de um mundo educacional interconectado, as possibilidade de atividades simultâneas e práticas no meio digital (MORAN; MASETTO; BEHRENS, 2000, p. 7) trouxeram para a área da educação – no final do século XX em diante – uma série de novas chances e expectativas para os seus objetivos pessoais de cada um em relação à educação. Isso recai diretamente ao fato de que a tecnologia passou a ser um intermédio de resolução de problemas e relação com espaço-tempo (TARCIA; CABRAL, 2012, p. 149).

No que concerne o ambiente de ensino e aprendizagem, focado na mediação por computador, muitos são os recursos utilizados para manter o interesse do aluno que vive inserido nesses meios em seu cotidiano, como a participação de fóruns, chats, posts, lista de e-mails, etc. Estas ferramentas, quando utilizadas de maneira efetiva, formam um ambiente de interação e mediação entre os alunos que estão inseridos no meio tecnológico a todo momento e transforma a maneira de ensinar, neste momento com o viés de interação e comunicação do círculo em que vivem (KENSKI, 2003, p.5).

É nesta proposta de mediação tecnológica e transformação do ambiente de ensino e aprendizagem, que se instaura o local em que esta pesquisa está situada: a EAD. A demanda por esta modalidade educacional passou por um aumento significativo, com destaque para a presente e crescente necessidade de disseminar informações em tempo real a nível mundial através dos meios digitais, principalmente no que diz respeito à troca de conhecimento específico de cada país (AZEVEDO, 2012, p. 4).

O aluno que será submetido a este tipo de aprendizagem poderá estar em qualquer lugar do mundo. Isto mostra que, além de ter esta flexibilidade, utilizando um dispositivo com *internet* para realizar um curso, o aluno também pode optar por uma área específica, de um determinado país, para buscar o conhecimento. No caso da disponibilização de um curso à distância para o mundo todo, vale ressaltar que há uma barreira comunicacional em relação aos idiomas em que serão distribuídas essas aulas, no caso desta pesquisa, no português brasileiro.

Essa pesquisa nasceu da necessidade do instituto de EAD Pecege[3] de se internacionalizar, justamente para se integrar nos movimentos atuais da sociedade. As aulas à distância oferecidas pelo Pecege são no formato de videoaulas ao vivo, com cursos de MBA voltados para diferentes áreas da Gestão. As aulas são oferecidas em português e, para que o instituto cumprisse o objetivo de ser internacionalizar, foi desenvolvida a ferramenta de inteligência artificial (IA) *Skylar*, de tradução automática-simultânea, com os idiomas português, inglês e espanhol, a fim de que as aulas fossem traduzidas automaticamente em formato de legendas durante as aulas oferecidas. O funcionamento da ferramenta consiste em 1) Captação da voz; 2) Transcrição automática; 4)

[3] O instituto será descrito na seção de Metodologia.

Segmentação das legendas; e 4) Tradução automática. Optou-se por averiguar apenas a transcrição devido ao volume que esta pesquisa geraria ao inserir os processos de tradução. O descritivo da ferramenta e sua estrutura estará exposto no Capítulo 3 – Metodologia.

Desta forma, essa pesquisa tem como objetivo o treinamento linguístico da IA contida no *software*, analisando os erros que ela gera durante o processamento das palavras a fim de se obter propostas de treinamentos para que ela gere conteúdos mais precisos. Busca-se compreender os parâmetros de legendagem em conjunção com o contexto em que os erros ocorrem nos segmentos apresentados. Embora a *Skylar* seja uma ferramenta de tradução, esta pesquisa tem seu foco apenas no processo de transcrição automática que a ferramenta realiza.

A análise dos erros e a proposta dos modelos linguísticos de treinamento serão estudadas e elaboradas utilizando os embasamentos teóricos, apresentado nos capítulos 1 e 2, sendo o Capítulo 1 - Linguística de *Corpus* (LC), que busca contextualizar, definir e apresentar algumas aplicações e métodos da área para a exploração dos corpora de forma quantitativa e qualitativa; e Capítulo 2 – Processamento de Linguagem Natural (PLN), que busca definir e contextualizar os métodos e aplicações em PLN ao longo do tempo, embasando o funcionamento do objeto da pesquisa.

O Capítulo 3 – Metodologia, se dedica a estabelecer os procedimentos dos experimentos. Nesta seção, serão descritos o *software* e seu funcionamento, o contexto em que ele se insere e os métodos que serão aplicados para a análise do *corpus* de transcrição. Para esta análise e a obtenção dos resultados que serão expostos no Capítulo 4 – Resultados e Discussão, os métodos foram divididos em duas partes, quantitativos e qualitativos. Para a análise quantitativa, se utiliza como base a compilação do corpus, a análise da quantidade de erros baseado no índice WER – *Word Error Rate* e a distância entre as correções que é apresentado pelo LD – *Levensthein Distance*. Para a análise qualitativa, foram modulados tipos de erros e a classificação dos segmentos por esses tipos, com o objetivo de entender quais tipos de erros ocorriam durante o corpus analisado.

De forma a alcançar os objetivos desta pesquisa, os seguintes questionamentos foram levantados:

1. Quais tipos de erros ocorrem durante a transcrição automática?

2. O contexto é relevante para que haja uma análise mais precisa do que precisa ser treinado na IA?

3. A estrutura do discurso do professor é representativa para que os segmentos sejam formados?

4. O formato de legendagem do *corpus* atrapalha o processamento automático?

A metodologia na qual esse trabalho se estabeleceu foi baseada na necessidade de se obter resultados direcionados para se responder às perguntas levantadas, pela eficiência das ferramentas de *corpus* ao se levantar resultados com uma grande quantidade de texto e das métricas da linguística computacional por contabilizar exatamente o que é necessário em termos de índice de erros.

Através dos métodos estabelecidos e na busca das respostas para as perguntas levantadas, esta pesquisa apresentou resultados quantitativos positivos quando apurados os índices de erros em até 49%, com uma amostra de 92% de precisão dos segmentos gerados. Os segmentos que obtiveram índice WER de 50 a 100% foram considerados negativos, representaram apenas 8% do recorte, tais informações.

Porém, este resultado quantitativo não apurava necessariamente quais tipos de erros ocorreram dentro dos segmentos analisados, dando motivação para a realização de uma análise qualitativa utilizando ferramentas de LC. Foram levantados os tipos de erros, sendo 1) Leve; 2) Médio e 3) Grave e foi realizada a distribuição dessas classificações nos segmentos. Neste momento da pesquisa, foram levantados os problemas específicos que a transcrição automática apresentava, fornecendo as respostas das perguntas de pesquisa e os insumos para os modelos linguísticos.

Por fim, o Capítulo 5 apresenta as Considerações Finais fazendo uma recapitulação do que foi analisado neste trabalho, elucidando a questão central, apresentando algumas limitações encontradas e as possibilidades de pesquisas futuras.

CAPÍTULO 1 - Linguística de *Corpus*

1. Introdução

Neste capítulo será apresentada a linguística de *corpus* (doravante LC), visando refletir sobre sua definição, por meio de um breve histórico e os métodos explorados para obtenção de dados linguísticos. O objetivo deste embasamento teórico é contextualizar a metodologia desta pesquisa, a qual é inspirada nos princípios de LC.

Na primeira seção, será exposta a linha do tempo da LC, procurando definir os conceitos e como a área se desenvolveu ao longo dos anos e suas implicações nas áreas correlatas. Na segunda seção, será discutida a composição de um *corpus,* a tipologia e as formas de análise possíveis na área. Na terceira seção, serão expostos os modelos estatísticos de análise de *corpus* e será discutida a abordagem metodológica na qual este trabalho se embasou.

1.1. Linguística de *Corpus* – Definição e histórico

De acordo com Berber Sardinha, 2004, podemos definir um *corpus* como:

> "(…) um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso 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." (BERBER SARDINHA, 2004, p.18)

Apesar de tal definição enfatizar que os dados devem estar em um formato que possa ser processado por computador, os princípios da LC datam da antiguidade. Nesse período, *corpus* era tratado em seu sentido original, corpo, e se tratava de um "conjunto de documentos (conforme o dicionário Aurélio)" (BERBER SARDINHA, 2004, p. 3). Ao longo do século XIII, o movimento por entender especificamente o que diziam os livros sagrados da Bíblia levou os estudiosos de tais textos a realizar um levantamento linguístico das passagens, palavras e as composições em que estas obras estavam inseridas. O objetivo era comprovar a divindade dos livros através da forma como estavam escritos (MCCARTHY e O'KEEFFE, 2010, p.3).

Na Grécia Antiga, há indícios da exploração de *corpus*. Um exemplo seria Alexandre, o Grande, que já realizava levantamentos linguísticos e culturais no *corpus* helenístico (BERBER SARDINHA, 2004, p. 3). A exploração de *corpora* ocorria com a utilização da prática de classificação de palavras e concordância, sendo o primeiro a listagem de ocorrências de um item específico para análise da palavra em contexto (SINCLAIR, 1991, p. 3) feita integralmente de forma manual com um propósito específico.

Estudos nessa linha são aprofundados com o trabalho de Becket que realizou análise de concordâncias de obras de Shakespeare. Seu objetivo era aprofundar o contexto e localização do autor para ampliar os conceitos aos pesquisadores, sendo que, para o estudo da literatura, suas pesquisas foram um recurso valioso, cujos trabalhos são vistos nas ferramentas de *software* dos dias atuais (MCCARTHY e O'KEEFFE, 2010, p.3; TRIBBLE, 2010, p.168). É possível afirmar que, a partir da década de 50, graças às pesquisas realizadas por estruturalistas americanos, que a possibilidade de fazer coleções pela captura de dados reais foi difundida, embora já existissem outras evidências da prática. Com isso, a prática de levantar dados linguísticos analisando-os de forma contextualizada, ou seja, dando significado ao conteúdo escrito/oral através de contexto foi difundida, dando ênfase ao ponto principal da pesquisa em LC (MCCARTHY e O'KEEFFE, 2010, p.4).

No final da década de 50, Chomsky apresentou a gramática gerativa. Uma teoria da linguagem que leva em conta estruturas sintáticas, semânticas e fonológicas como uma forma padronizada de se expressar, contrariando as coleções de dados linguísticos empíricos como a construção de um *corpus*. Essa ideia fez com que os estudos da linguagem se afastassem do empirismo justamente por demonstrar uma padronização baseada nas estruturas da linguagem (BERBER SARDINHA, 2004, p. 4).

Não somente, a teoria chomskyana pôs em prova o desenvolvimento de pesquisas em LC, mas também a prática de análise de *corpus* sofria mudanças e passava a demandar um método que não fosse impactado com a margem de erro derivada da análise e organização manual dos dados (BERBER SARDINHA, 2000, p. 325). Porém, como ainda não existia uma tecnologia que pudesse abranger este novo método, a área sofreu outras mudanças significativas em seu *modus operandi.* A área da LC vinha enfrentando algumas críticas a respeito do processo manual de coleta dos dados, pois mesmo com

muitos analistas trabalhando, havia um grande volume de inconsistência na análise. Esses fatores contribuíram para que a área passasse por uma queda no interesse em se continuar desenvolvendo pesquisas nesse âmbito (BERBER SARDINHA, 2004, p. 4).

Até meados do século XX, o processo de construção e análise de *corpus* era limitado, por ser muito dispendioso e manual. Dado o volume e a profundidade dos dados que são analisados na LC, tais métodos se revelaram ineficientes na prática de análise do *corpus*, passando a demandar uma maior eficiência para lidar com a margem de erro derivada da análise e da organização manual dos dados, conforme exposto anteriormente (BERBER SARDINHA, 2000; 2004; SINCLAIR, 1991). No final dos anos 50, a invenção dos gravadores portáteis foi de extrema importância para os estudos empíricos da linguagem. A transcrição de mais de 166.000 palavras em inglês, extraídas de conversas informais, foram compiladas no primeiro *corpus* oral na Universidade de Edimburgo por uma iniciativa de Sinclair, de 1963 a 1965 (BONELLI, 2010, p. 16).

Então, com o surgimento dos *mainframes*[4], nos anos 60, as pesquisas da linguagem tiveram um aliado no processamento de dados e ganharam força novamente. Em 1961, a apresentação do primeiro *corpus* eletrônico, o *Brown Corpus*, foi realizado por Francis & Kucera (1967), com 1 milhão de palavras de documentos coletados no mesmo ano. Tal feito era tido como inimaginável até poucos anos atrás e tornou-se realidade por meio da pela revolução tecnológica (BONELLI, 2010, p. 15). O *Brown Corpus* foi um importante marco para a área pois, pela primeira vez, havia sido processado um *corpus* robusto o suficiente possível de ser analisado sob a ótica computacional. Através de elementos interdisciplinares, a compilação e análise do *Brown Corpus* resultou em um avanço para a área de linguística computacional (MCENERY; HARDIE, 2011, p. 99).

Além de um melhor processamento tecnológico dos dados, a década de 60 foi crucial não somente para armazenamento de textos escritos, mas também de conteúdos falados. Sinclair foi o precursor dos estudos do léxico com *corpus* falado (SINCLAIR, 1997; 1966) e, embora o *Brown Corpus* e o *corpus* falado desenvolvido na Universidade de Edimburgo fossem desconhecidas um do outro, ambos foram processados na mesma

[4] O *mainframe* é um computador de grande porte que foi um dos primeiros hardwares com um processamento robusto de dados. Hoje ainda é utilizado para grandes processamentos que requerem maior segurança e potência (ABBATE, 1999, p. 1).

época, graças a evolução do desenvolvimento tecnológico nas áreas da linguagem (BONELLI, 2010, p. 16).

É notável a relação entre o desenvolvimento dos computadores e a LC. No final dos anos 60, na transição entre os *mainframes* e os microcomputadores pessoais – os quais eclodiram mais especificamente nos anos 70, já era possível visualizar grandes máquinas de calcular com operações estritamente quantitativas e mecânicas (ABBATE, 1999, p. 1695). O fato é que, para o processamento estatístico de grandes dados, o computador apresentava um resultado rápido e preciso. Porém, para a análise qualitativa, o humano fazia essa atividade consideravelmente de forma mais lenta, só que com uma precisão muito maior (MCENERY; HARDIE, 2011, p. 228)

Os anos de 1970 demonstraram um cenário de consolidação da pesquisa computadorizada em LC, colocando em prática o que havia sido desenvolvido na década anterior. A tecnologia disponível promoveu uma quebra de paradigma para a área, ainda que escassa e lenta, com pouca inteligência[5] no processo e sem recursos de *software* que manipulassem os dados linguísticos de forma concisa. Mesmo com este percalço no avanço de processamento com e ferramentas que pudessem efetivamente ajudar a alavancar a área da LC, a década de 70 proporcionou alguns marcos como: (i) mais que 1 milhão de palavras processadas; (ii) proporcionar anotação em *corpus* e a (iii) possibilidade de transcrição fonológica para objetos verbais (BONELLI, 2010, p. 16).

A disponibilidade de sistemas especializados em exploração de *corpora* se expandiu nesta década, pois além do avanço constante em termos de armazenamento e processamento, algumas ferramentas de análise foram desenvolvidas. Um dos exemplos é o TAGGIT, o primeiro etiquetador morfossintático[6] em 1970, que foi responsável por etiquetar 77% do *Brown Corpus* automaticamente, sendo os 33% restantes feitos de forma manual. Outro exemplo é o CLAWS, de 1979, proveniente do TAGGIT, para etiquetar os *LOB Corpus*, iniciativa semelhante ao *Brown* (referente ao inglês britânico). O CLAWS nasceu com a intenção de aprimorar a análise já realizada pelo TAGGIT, na

---

[5] "*Computers were still calculating machines with small memories, and programming languages were not devised with the manipulation of character strings in mind*" (BONELLI, 2010, p. 16).

[6] O processo de etiquetagem consiste na marcação do texto com informações linguísticas relevantes para o que ser quer pesquisa, ou seja, a etiquetagem serve como um método que padroniza as estruturas linguísticas. Pode ser feito de forma manual ou através de *software*s (etiquetadores) que fazem a etiquetagem de forma automática (BIBER; CONRAD; REPPEN, 1998, p. 257).

aplicação de métodos estatísticos de análise (BERBER SARDINHA, 2004, p. 15; LÜDELING; KUTÖ, 2008, p. 33).

Pode se dizer que a história dos computadores e da LC andaram e se desenvolveram juntas. À medida que havia mais poder de processamento, havia mais possibilidades de manipulação dos dados depositados no processo de análise linguística, propiciando uma análise estatística mais consistente para o processo. Os computadores mudaram a forma como os linguistas lidavam com os *corpora* (KENNEDY, 1998, p. 4). A relação entre a LC e a era da revolução na tecnologia, propiciada pelo uso dos computadores desde os anos 60, passou por três fases importantes. A primeira fase se destaca pela a utilização do computador para processar e armazenar dados; a segunda, além do processamento e armazenamento de uma maior quantidade de dados, possibilita maior margem para a exploração de novos métodos quantitativos de análise de *corpus*. A terceira, já entrando na década de 90, proporcionou a extrapolação da pesquisa quantitativa em *corpus*, promovendo o desenvolvimento qualitativo significativo e modificando completamente a forma de se trabalhar com estes dados linguísticos (KENNEDY, 1998, p. 5; BONELLI, 2010, p. 17).

Nos anos 80, cerca de 20 anos após o primeiro contato dos *corpora* com tecnologias de processamento através do computador, os microcomputadores pessoais revolucionaram novamente a área, com melhores memórias, processamento e desenvolvimento de *software* moldados com base na organização manual feita anteriormente, formulando as diretrizes da pesquisa computadorizada (BERBER SARDINHA, 2000, p. 238; KENNEDY, 1998, p. 4). A partir da década de 1990, o papel dos computadores passou a ser fundamental em pesquisas linguísticas. A forma de lidar com os dados quantitativos, a precisão em que os processos estatísticos eram apresentados pela máquina e a disposição dos dados de forma rápida foram pontos que facilitaram as bases de *corpus* para estudos como processamento de linguagem natural (doravante PLN), síntese de voz, tradução automática, análise de contexto e descrição de linguagem para o ensino de idiomas (KENNEDY, 1998, p. 6; ALUÍSIO & ALMEIDA, 2006).

As questões que giram em torno da área da LC, portanto, não se mantêm apenas na manipulação de dados através de *software*s ou gerando dados probabilísticos para exploração do *corpora* (HALLIDAY, 2005, p. 63), mas também buscam responder questões teóricas e aplicadas em relação à linguagem (KENNEDY, 1998, p. 3). Halliday

(2005), discute a história da LC e o papel do computador nas pesquisas da linguagem. Para o autor, a ampliação do escopo da LC que o computador e as ferramentas computacionais de processamento linguístico progrediram na maneira de relacionar as palavras com seu sentido, ou seja, saíram da posição de dados estatísticos e passaram a se desenvolver de forma qualitativa (HALLIDAY, 2005, p. 242). Assim, com o uso das ferramentas computacionais para a análise de *corpus*, é possível realizar observações envolvendo morfologia, sintaxe, semântica, discurso, entre outros. Saindo do modo processual quantitativo dos primórdios (HALLIDAY, 2005, p. 241), a pesquisa passou a percorrer atividades como a produtividade e o emprego de palavras, expressões e gramática, além de tornar-se base para a descoberta de novos fatos linguísticos, gerados nestas ferramentas, como discutiremos nas seções a seguir (BERBER SARDINHA, 2000; ALUÍSIO & ALMEIDA, 2006).

Para compreender os estudos da linguagem baseados em *corpus*, há uma divisão entre análise de estrutura e análise de uso que precisa ser considerada (BIBER; CONRAD; REPPEN, 1998, p. 135). Quanto à análise de estrutura, podem ser observadas as identificações estruturais da linguagem, como: morfemas, palavras, frases, classes gramaticais e composições gramaticais específicas. Já a análise de uso, trata-se da investigação de como os falantes de uma língua exploram seus recursos linguísticos, através de reconhecimento de padrões e fatores contextuais que influenciam em diversos tipos de variação, como as de gênero e registro (BIBER, 1995; BIBER; CONRAD; REPPEN, 1998).

Isto posto, acerca do histórico da LC, seguem-se, nas próximas seções, tentativas de esclarecer de forma técnica a formação de um *corpus*, processos metodológicos quantitativos e qualitativos de análise e como as ferramentas computacionais trabalham para um maior aproveitamento dos resultados obtidos.

1.2. Conceitos, caracterização e composição de um *corpus*

Esta seção se dedica a apresentar a caracterização de um *corpus*, os conceitos básicos de montagem e coleta, exemplos de processos metodológicos aplicados e as ferramentas em que os dados podem ser processados. Como conclusão, esta seção exibe características específicas de *corpus* para transcrição, tema central deste estudo.

Há diversas possibilidades de construção de *corpora*, sendo que sua composição está relacionada aos objetivos de pesquisa (LÜDELING; KUTÖ, 2008, p. ix). Estes mesmos objetivos podem ser determinantes ferramentas a ser utilizadas, dada sua variedade e possibilidade de aplicação.

Devido ao caráter interdisciplinar que as pesquisas em LC apresentam, além de se estabelecer metodologias únicas para cada caso, há também uma série de considerações na compilação de um *corpus* (MCENERY; HARDIE, 2011, p. 11). Todavia, como apresentado por Esimaje e Hunston (2019), o escopo de compilação para um *corpus* deve seguir, pelo menos, três características principais, sendo: 1) textos que ocorrem de forma natural; 2) textos selecionados para exploração de variedade ou registro linguístico específico; e 3) textos com grande volume que necessitem ser processados e armazenados em máquinas (ESIMAJE; HUNSTON, 2019, p. 7).

Berber Sardinha (2004) define a construção de um *corpus* a partir de sua relação a diversos fatores característicos em meio aos quais esta exerça uma representatividade independentemente do método em que será desenvolvida a pesquisa (BERBER SARDINHA, 2004, p. 20). Complementando a discussão de Esimaje e Hunston (2019), a compilação de um *corpus* deve seguir alguns pré-requisitos como: a origem, o propósito, a composição, a formatação, a representatividade e a extensão (BERBER SARDINHA, 2004, p. 18).

A origem está relacionada com a autenticidade do *corpus*. Por autenticidade, entende-se que a formação do texto utilizado para pesquisa não pode ter sido feita intencionalmente, ou seja, deve ser coletado ou gerado através de fontes nas quais foi naturalmente produzido. Um exemplo são as transcrições de conversas ou outros conteúdos falados, foco central desta pesquisa. Além de textos produzidos em *blogs*, jornais, *sites* de redes sociais, etc. O propósito, composição e formatação, por sua vez,

são itens que recaem sobre a escolha criteriosa dos textos que vão compor o *corpus*. Enfatizando ainda a autenticidade, o *corpus* deve ser formatado de forma condizente com as diretrizes do que se quer explorar, evitando a inconstância nos resultados. Por fim, a representatividade deve ser imposta respeitando o objetivo da pesquisa e deve ser tratada separadamente em cada caso. Além disso, está diretamente ligada à extensão do *corpus*, uma vez que a função representativa está relacionada com a dimensão dele (BERBER SARDINHA, 2004, p. 19–20; ESIMAJE; HUNSTON, 2019, p. 16).

É importante ressaltar que caracterização do *corpus* e de sua composição são os itens que guiam a construção de um objeto de pesquisa legítimo, cujo irá dar respostas para novas descobertas na linguagem (REPPEN, 2010, p. 31). É essencial levantar as perguntas bem claras e direcionadas para que haja uma caracterização específica do que se irá coletar, justamente pela função representativa do *corpus* vir a ser o fator que determinará sua dimensão (BERBER SARDINHA, 2004, p. 19).

Biber (1993) discorre a respeito da elaboração do *corpus* através de ciclos, iniciando pela escolha da questão de pesquisa, dos gêneros textuais que serão envolvidos, das análises empíricas da língua e de suas variações. Por fim, considera a revisão e a execução com as metodologias específicas para responder às questões elaboradas inicialmente. Portanto, a ênfase sempre deve ser por meio da sua representatividade (BIBER, 1993, p. 1).

Uma vez que as perguntas de pesquisa estão elaboradas, pode-se começar a coletar o material e a produzir o *corpus* (REPPEN, 2010, p. 32). Visto isto, Berber Sardinha (2004) discute a tipologia do *corpus* e quais as perguntas a serem levantadas para cada item, determinando sob qual investigação o *corpus* deverá ser delineado. Esta composição deve ser criteriosa e é apresentada por meio de um caráter técnico, ou seja, considerando procedimentos bem, para que haja a correspondência no que se procura resolver (BERBER SARDINHA, 2004, p. 19; LÜDELING.; KUTÖ, 2008, p. ix).

São levantados grupos de critérios para a construção do *corpus* e esses grupos são, posteriormente, bases para a classificação do *corpus*. São sete grupos provenientes de vários autores: modo, escrito ou falado; tempo, sincrônico, diacrônico, contemporâneo ou histórico; seleção: amostragem, monitor, dinâmico, estático ou equilibrado; conteúdo, especializado, regional ou multilíngue; autoria, de aprendiz ou de língua nativa;

disposição interna, paralelo ou alinhado; finalidade, de estudo, de referência, de treinamento ou teste (BERBER SARDINHA, 2004, p. 19).

No que se refere ao modo, a construção do *corpus* pode ser direcionada à coleta de trechos escritos ou falados, que ocorram de forma natural, direcionando-se para a representatividade do que se pretende investigar (BERBER SARDINHA, 2004, p. 28; REPPEN, 2010, p. 33). Para coletânea de dados escritos, busca-se o propósito da pesquisa e depois a coleta nos locais onde podem ocorrer o tema pesquisado. No caso do *corpus* falado, base desta pesquisa, buscam-se as fontes e há a etapa de transcrição desses dados antes do uso (REPPEN, 2010, p. 33).

Na classificação de tempo, o *corpus* pode ser sincrônico, diacrônico, contemporâneo ou histórico. O sincrônico é relativo a apenas um período. O diacrônico, por sua vez, refere-se a um conjunto de variações ao longo de diferentes períodos de tempo, contemporâneo para o que ainda está ocorrendo e passado para fatos que já ocorreram (BERBER SARDINHA, 2000, p. 340, 2004, p. 20).

A seleção de um *corpus* é dividida entre amostragem, monitor, dinâmico, estático ou equilibrado. O *corpus* de amostragem separa uma parcela finita de textos compilados para se obter o máximo de respostas a respeito de uma questão relacionada a língua. De outro modo, o monitor procura investigar como a língua se apresenta na atualidade. Os tipos dinâmico e estático dão suporte e caracterizam o *corpus* de amostragem e o monitor, sendo o primeiro ajustável quanto ao tamanho e o segundo sem condições de modificação no que foi coletado, respectivamente. No equilibrado, a distribuição entre gêneros textuais é balanceada em quantidades equivalentes (BERBER SARDINHA, 2000).

O conteúdo pode ser especializado, regional ou multilíngue, sendo (i) o especializado um conjunto de textos com registros específicos, (ii) o regional, com variedades linguísticas características de lugares diferentes e (iii) o multilíngue, a inserção de diversos idiomas diferentes. No que diz respeito à autoria, esta pode ser de (i) aprendiz, quando as ocorrências não são provenientes da língua nativa e de língua nativa, que visa analisar fenômenos provenientes da língua nativa (BERBER SARDINHA, 2000).

A disposição interna diz respeito à forma como esse *corpus* será analisado. Pode ser (i) paralela, um formato que permite a comparação entre o original e tradução (ou o

original e o ideal, no caso exposto no método dessa pesquisa) ou (ii) alinhada, formato no qual segmentos se dispõem embaixo um do outro (BERBER SARDINHA, 2000).

A finalidade de utilização do *corpus*, que pode ser de estudo, de referência, de treinamento ou teste. Este critério é um dos principais a ser considerado na compilação do *corpus*, pois se relaciona diretamente com o problema de pesquisa levantado, determinando seu formato e construção. O *corpus* de estudo é descritivo quando são estudadas questões particulares para a pesquisa em questão, ou de referência, quando usado de forma comparativa ao *corpus* de estudo e o de treinamento ou teste é específico para uma aplicação de desenvolvimento de ferramentas ou treinamento estipulado. (BERBER SARDINHA, 2000).

Além do teor técnico dos requisitos apresentados, há uma etapa adicional que diz respeito aos direitos do conteúdo que estão sendo utilizado para a formação do *corpus*, os quais incluem seguir "regras legais para obtenção de direitos de uso do material junto a autores e editores que detêm o *copyright* do texto ou consentimento de indivíduos cujos direitos de privacidade devem ser reconhecidos" (ALUÍSIO; ALMEIDA, 2006, p. 160).

Feita essa discussão, a próxima seção se debruça nos aspectos metodológicos de análise de um *corpus* e alguns exemplos de aplicações estatísticas possíveis na área da LC, direcionando-se para a metodologia de análise desenvolvida neste estudo.

1.3. Métodos e aplicações de análise em *corpus*

Halliday (2005) discute a linguagem como um sistema probabilístico (HALLIDAY, 2005, p. 45). O autor discorre a respeito da probabilidade e de como a linguagem pode ser analisada por este viés, considerando não só a gramática[7], mas também interpretando a escolha em que essas palavras foram inseridas nos objetos de análise em questão. Embora a busca por uma padronização nas ocorrências lexicais de uma língua tenha sido discutida por muitos autores (BERBER SARDINHA, 2004, p. 39), Firth (1957) elucida que o sistema probabilístico está ligado ao contexto em que as palavras se encontram (BERBER SARDINHA, 2000, p. 360, 2004, p. 41; BIBER; REPPEN, 2015, p. 16; LÜDELING; KUTÖ, 2008, p. 351) e explica com sua famosa afirmação "*you shall know a word by the company it keeps*" (FIRTH, 1957).

A visão da linguagem como um sistema probabilístico surge com a introdução de alguns conceitos levantados por Firth (1957) e a partir disso que nascem os métodos de análises computacionais dos *corpora*. Firth propôs que o significado não estava simplesmente no léxico, mas em todo o contexto situacional analisado de forma integrada, além das condições gramaticais e semânticas em que o trecho se encontra (BERBER SARDINHA, 2004, p. 41). Um exemplo de conceito introduzido por Firth é a colocação.

A colocação tem por objetivo fazer a associação entre o léxico e a semântica, ou seja, analisar a ocorrência de um termo através do seu significado e das palavras que estão ao redor desse termo ou no decorrer do texto (SINCLAIR, 1991, p. 54). Por exemplo, para o desenvolvimento da linguística computacional, usa-se a colocação para treinamento de sistemas inteligentes com a intenção de analisar além do léxico, elaborando um sistema de inferência interpretativa mais preciso (LÜDELING, A.; KUTÖ, 2008, p. 351).

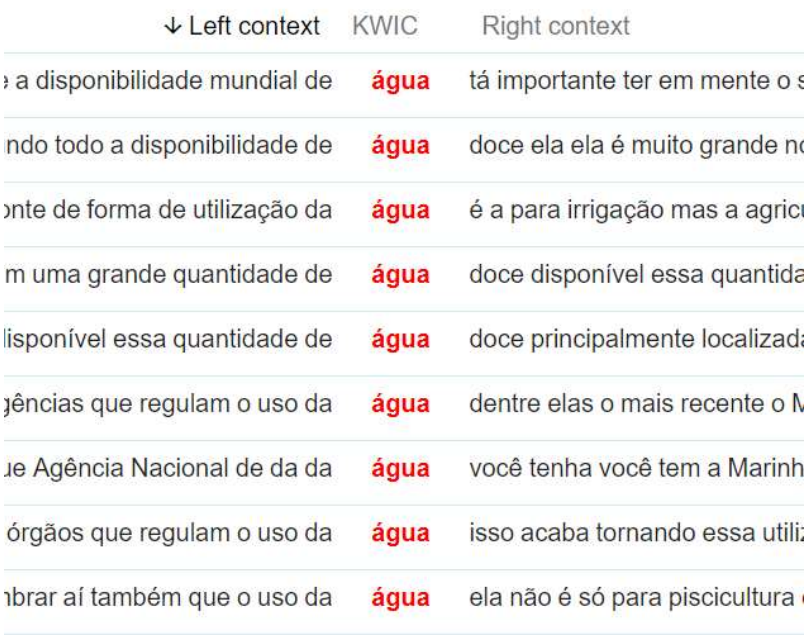
A partir das listas de palavras que são levantadas para demonstrações estatísticas simples (SINCLAIR, 1991, p. 32), e da colocação citada acima, também podemos analisar um *corpus* executando uma concordância. Concordância é um compilado de ocorrências de determinado léxico e seus contextos (SINCLAIR, 1991, p. 32), ou seja, isola-se o termo específico que deve ser explorado e se tem o contexto das palavras que

---

[7] Para Halliday (2005), o termo "gramática" (do inglês, *grammatics*), significa que a construção estrutural expressa um significado de linguagem (HALLIDAY, 2005, p. 45).

ocorrem ao seu redor. A análise pode ser feita por meio da extração das palavras mais frequentes, isolando os termos específicos no formato KWIC (*Key word in context*), de forma a centralizar as palavras e seus contextos de forma rastreável[8]. A Figura 1, abaixo, demonstra um modelo de concordância executada no *software Sketch Engine*[9].

Figura 1 - Exemplo de concordância



↓ Left context	KWIC	Right context
a disponibilidade mundial de	**água**	tá importante ter em mente o s
ndo todo a disponibilidade de	**água**	doce ela ela é muito grande nc
nte de forma de utilização da	**água**	é a para irrigação mas a agricı
m uma grande quantidade de	**água**	doce disponível essa quantida
isponível essa quantidade de	**água**	doce principalmente localizada
gências que regulam o uso da	**água**	dentre elas o mais recente o M
ue Agência Nacional de da da	**água**	você tenha você tem a Marinh
órgãos que regulam o uso da	**água**	isso acaba tornando essa utiliz
nbrar aí também que o uso da	**água**	ela não é só para piscicultura

Fonte: Captura de tela do *software Sketch Engine.*

No exemplo acima, a Figura 1 demonstra a concordância executada com o termo isolado "água", em busca do contexto em que esse substantivo ocorreu durante o *corpus* em questão. Pode-se analisar essa concordância através de seus colocados, diminuindo ou aumentando as possibilidades que estão ao redor do termo isolado. Em todas as ferramentas que executam concordância é possível pesquisar, através das palavras mais frequentes, de diversas formas em que o termo aparece, de forma a responder às perguntas de pesquisa em que está sendo executada (RAYSON, 2015, p. 33).

As análises são feitas atualmente utilizando ferramentas computacionais que aumentaram, ao longo dos anos, as possibilidades de pesquisas baseadas na LC (RAYSON, 2015, p. 33). O desenvolvimento de metodologias acerca de LC não se

[8] Os *softwares* de LC permitem que se busque os termos isolados no KWIC para a busca dos seus trechos específicos.
[9] https://sketchengine.co.uk

mantém apenas para a busca de padrões através de expressões, mas também para se obter os registros daquele *corpus* em contextos situacionais (BIBER; CONRAD; REPPEN, 1998, p. 135).

Biber (2010) discorre a respeito da perspectiva de registro e da perspectiva de gênero como duas formas distintas de se enxergar e analisar o corpus. Além disso, considera importância de organizar o texto de forma estruturada em categorias (BIBER, 2010, p. 241). Na perspectiva de gênero, deve ser analisa a estrutura linguística em textos completos, ou seja, analisar a construção do texto de forma estrutural. Já na visão de registro, o foco se concentra nas características representativas do texto e na variação entre essas características, buscando analisar formas de uso em contexto. Desta forma, os estudos baseados em corpus geralmente elucidam o registro, ao invés do gênero do texto, pois busca investigar a variação do uso das estruturas linguísticas, e não as características do texto em si (BIBER, 2010, p. 242).

É importante ressaltar que, para o propósito desse estudo, é relevante entender em que contexto o corpus elaborado se refere. Nesse caso, como discorre Biber et. al. (1998) ainda a respeito da variação de gênero e registro, os padrões de ocorrência variam no caso de um corpus falado e um escrito (BIBER; CONRAD; REPPEN, 1998, p. 143). Williams (2002) também discorre a respeito da criação de critérios para análise de corpus falado, de forma a cumprir o objetivo de exploração dentro de aplicações de processamento de linguagem natural (PLN).

Por isso, o próximo capítulo se destina a explorar teoricamente as premissas do PLN, buscando ressaltar quais pontos são necessários para a compreensão dessas ferramentas, a fim de se elaborar uma pesquisa utilizando a compilação de um corpus falando originário de um software de PLN.

CAPÍTULO 2 – PROCESSAMENTO DE LINGUAGEM NATURAL

2. Introdução

Neste capítulo, será apresentado o processamento de linguagem natural (doravante PLN), visando apresentar sua definição, um breve histórico acerca de suas aplicações em termos computacionais e contribuições para a área da sintetização da fala e da tradução.

Na primeira seção, será exposta a definição e um breve histórico sobre o PLN. Na segunda seção, serão apresentadas as aplicações em PLN e quais são suas possibilidades, na tentativa de traçar percurso evolutivo e histórico. Na terceira seção, se discorrerá à respeito de tradução e tecnologias, pois o PLN se desenvolveu fortemente nessa área e na área de reconhecimento e sintetização da voz. E, por fim, será destacado o movimento da pós-edição de conteúdos gerados automaticamente através de ferramentas de PLN.

2.1. PLN: Definição e histórico

O PLN é a área da computação que se preocupa com a relação entre a máquina e as línguas naturais, ou seja, a área que lida com o tratamento linguístico do processamento da língua pela máquina. Nesse sentido, o propósito dos estudos em PLN refere-se à construção de *software*s envolvendo tradutores automáticos, *chatterbots*[10], *parsers*[11], reconhecedores automáticos de voz, entre outras aplicações com propósitos interpretativos da linguagem (OTHERO, 2006). O PLN pode ser definido, então, como o conjunto de técnicas e algoritmos computacionais que realizam a análise linguística de textos, bem como fazem sua representação para realizar tarefas com o processamento da linguagem humana (LIDDY, 2001).

A respeito do desenvolvimento de aplicações utilizando PLN, muitos subsistemas e recursos computacionais são utilizados, por conta de aspectos estruturais da língua como morfologia, sintaxe e semântica (OTHERO, 2006). Assim, o PLN engloba diversas áreas para a exploração de sistemas linguísticos. Desta forma, podemos traçar um paralelo entre os subsistemas utilizados para realização de softwares de PLN com a área da LC.

---

[10] "For a regular definition, a chatbot is a specific kind of robot, that is, an automated agent designed to execute repetitive tasks that are deemed useful for someone (who, herself, can either be a human or another bot)" (DORR; SNOVER; MADNAMI, 2010).

[11] *Parsers* são *software*s de análise de entradas sintáticas de forma a analisar estruturas gramaticais (DORR; SNOVER; MADNAMI, 2010).

Conforme apresentado no capítulo 1, a área da LC se ocupa, em uma de suas vertentes, da exploração de *corpora* para análises de conjuntos linguísticos de forma estatística (BERBER SARDINHA, 2004). Desta maneira, podemos dizer que a contribuição da LC para sistemas de PLN torna-se primordial e relevante para o desenvolvimento de sistemas voltados para linguagem.

Historicamente, as pesquisas em PLN datam do final da década de 1940, com a criação de uma máquina de tradução em que foi utilizado o uso de teorias de criptografia e teoria da informação para realizar tradução de linguagens. Porém, esses modelos básicos de tradução produziam resultados não satisfatórios ao desconsiderar conceitos básicos da linguagem natural, como a ambiguidade léxica. Neste, a linguagem estava sendo desenvolvida através de criptografia sem levar em consideração o teor gramatical (LIDDY, 2001, p. 3). A partir da década de 50, foram se desenvolvendo sistemas voltados à precaução dos erros gramaticais produzidos pelas ferramentas iniciais, buscando um melhor entendimento na produção de melhores máquinas de tradução com o uso da linguística. Nesta mesma década, outros campos de PLN começaram a emergir, como o reconhecimento de voz (LIDDY, 2001, p. 5). Foi neste movimento que o primeiro sistema de reconhecimento de voz foi criado na *Bell Labs*, em 1952: "Audrey", uma máquina capaz de reconhecer dígitos numéricos, entre outras atividades. Com isso, avanços nessa área começaram a emergir, como na Feira Mundial de 1962, quando a IBM apresentou o "Shoebox" que reconhecia 16 palavras da língua inglesa (LIDDY, 2001).

Avanços computacionais e em pesquisas permitiram muitos progressos na área de reconhecimento de voz, sendo a década de 80 marcante pelo uso do modelo estatístico *Hidden Markov Model*. Este modelo permitia o uso de probabilidades para determinar sons como palavras (JUANG; RABINER, 2004, p. 12). A década de 80 foi decisiva para o futuro dos sistemas de PLN, que até então se baseavam em um sistema de regras fixas utilizando teorias da gramática gerativa, levando apenas a sintaxe em consideração na interpretação das frases. Nesta década, se iniciaram os estudos em aprendizado de máquina para aplicações em PLN, implementando análises de LC a fim de elaborar um conteúdo automático baseado não somente em estrutura, mas também em semântica (DASCAL, 2014; KHUMMONGKOL; YOKOTA, 2015).

As aplicações em PLN começaram a caminhar para um caminho mais complexo em relação a compreensão das palavras, passando a envolver aplicações que utilizam

abordagem cognitiva e intuitiva dentro do campo da linguagem (DASCAL, 2014). A esse campo, atribui-se o nome de Compreensão de Linguagem Natural[12]. Este visa não apenas processar as linguagens naturais, como amplamente discutido em PLN, mas conceber critérios a ponto de proporcionar o avanço das máquinas nos campos, principalmente, semânticos e pragmáticos, os quais se mantêm, teoricamente, na cognição apenas de seres humanos (HUTCHINS, 2007; OTHERO, 2006). Nas próximas décadas após os anos 80, com o uso comercial de reconhecimento de voz em brinquedos e também em empresas, como a *Dragon Dictate*[13], por exemplo, foram possíveis aprimoramentos nos algoritmos de reconhecimento, que elevaram a capacidade de entendimento das máquinas, principalmente com o advento de técnicas avançadas de Inteligência Artificial, isto é, o *Deep Learning* (ABDEL-HAMID et al, 2014).

Contudo, apesar de tecnologias nas áreas da linguagem propiciarem melhores resultados de precisão no reconhecimento de voz, muitas ainda não atingem 100% de precisão quando reconhecendo diversos sons. Essa variação nos resultados pode se dar pela má qualidade do equipamento que capta a voz, influência de sons externos, nuances na fala do locutor, entre outros aspectos (RANI; GIRIJA, 2012). Esta pesquisa busca melhorar a precisão do sistema de reconhecimento de voz, identificando modelos linguísticos permitindo uma melhor identificação das falas e produção dos resultados de voz para texto.

A linguística atua com interesses complementares aos da computação dentro do PLN, provendo as estruturas necessárias para o entendimento da linguagem natural que sistemas computacionais necessitam para realizar a representação e análise de linguagens (BRISCOE, 2013). Conforme supracitado, o PLN é uma área interdisciplinar que necessita de diferentes perspectivas para que tenha êxito. Assim, o PLN e a Linguística podem ser vistos como congruentes já que podem ser utilizadas em conjunto. Finatto et. al (2015) discorrem que o fator diferenciador aqui está na análise estatística da língua em uso efetuada em LC, enquanto o PLN busca (com o auxílio de *corpus*) criar soluções

---

[12] *NLU – Natural Language Understanding* (KHUMMONGKOL; YOKOTA, 2015).
[13] *Dragon Speech Recogniton* – é uma ferramenta proveniente do pacote de reconhecimento de voz da Dragon Naturally Speaking, produzido pela *Nuance Communications*.

aplicáveis a problemas pontuais que utilizem linguagem em sua especificação (CIULLA; LOPES; FINATTO, 2015).

O PLN executa diversos tipos de processamentos em todas suas aplicações, como processamentos sintáticos, semânticos e pragmáticos. Entretanto, dentro dos níveis de análise linguística, o morfológico e o lexical são as mais facilmente abstraídas por sistemas computacionais devido às suas características estruturais exatas. Já os níveis sintático, semântico e pragmático requerem, devido à sua complexidade, modelos linguísticos a serem identificados e aplicados com o objetivo de representação e análise de processamentos dentro de toda a área de PLN (BRITO, 2000).

O processamento sintático desempenha o papel de especificar as estruturas gramaticais de uma língua, formalizando assim a análise morfológica e lexical, utilizando do formalismo gramatical e de um analisador (*parser*) para processar e montar a sentença na ordem estabelecida da língua em questão (BRITO, 2000). Nesse contexto, o *parser* (NILSSON et al., 2009) é a estrutura responsável por realizar a identificação do significado exato de palavras, ou seja, a identificação de quais palavras foram reconhecidas pelo *software*, que será necessário juntamente com os formalismos gramaticais por efetivar a transmissão para o processamento semântico.

O processamento semântico leva em consideração os "elementos de significação" (BRITO, 2000, p. 14), os quais contemplam a estrutura do processamento sintático. Brito (2000) ainda comenta que há uma vasta discussão a respeito desses dois itens e de suas relações na linguística computacional, levantando questões importantes a serem delineadas em um sistema de PLN. A formulação de um processamento de linguagem não deve ser somente sintático, ou seja, levando em conta somente a estrutura sintática dos elementos e, sim, processado com elementos de significação para se realizar "o real objetivo do entendimento da linguagem natural em computadores" (BRITO, 2000, p. 14–15).

Há muitos outros aspectos a serem discutidos acerca do PLN e sua estrutura. Porém, este estudo se dedica à fundamentação introdutória e às aplicações de PLN para basear a investigação aqui realizada. Na próxima seção, se discorrerá a respeito das aplicações específicas de PLN.

2.2. Aplicações em PLN

Esta subseção se dedica a introduzir tipos diferentes de aplicação de PLN, bem como explorar a área que mais se utiliza deste processamento, a tradução automática. Os desenvolvimentos de aplicações em PLN são interdisciplinares, focando principalmente em alguns tipos de aplicações como: (i) as máquinas de tradução, (ii) a construção de linguagem para sistemas robotizados (geração de linguagem natural), (iii) reconhecimento de voz, (iv) compreensão de linguagem natural, (v) reconhecimento óptico de caracteres (OCR), extração de dados através de análises em linguística de *corpus* (*big data*), entre outras aplicações (CRACIUNESCU; SALAS; KEEFFE, 2004; FIEDERER; O'BRIEN, 2009; HUTCHINS, 1995).

As aplicações em PLN buscam auxiliar na solução de determinadas aplicações no campo da linguagem, fazendo com que o *software* interaja em diferentes aspectos com o usuário. Com o aprendizado de máquina no desenvolvimento das aplicações em PLN, os processos práticos acerca da interação entre homem e máquina se torna facilitado, envolvendo também o envio de parâmetros técnicos para o aprimoramento de cada *software* com estes recursos.

Com a evolução da tecnologia nos campos da linguagem, os tradutores automáticos tomaram a frente devido à necessidade desse recurso em diversas áreas de atuação. Nesse caso, sabemos que ferramentas com esse auxílio transformam o trabalho da tradução de forma a atender mais demandas, porém também sabemos que a eficácia do tradutor automático não realiza uma tradução precisa o suficiente, por diversos motivos científicos que ainda não permitem uma tecnologia que replique o processo natural de voz de uma forma coerente (OTHERO, 2006).

Apesar de sabermos que a tradução automatizada nunca será tão eficaz quanto a humana, devido "à complexidade e riqueza das línguas humanas, que relutam em se entregar à formalização do computador" (OTHERO, 2006), o processo automático desenvolvido em uma ferramenta com IA, sendo que esta ferramenta aprende com os erros conforme alterada, é uma saída para a otimização para os equívocos ocorridos pelo processo inteiramente automático.

2.3. Tradução e tecnologias

Esta subseção expõe o histórico acerca da tradução automática e suas vertentes dentro do desenvolvimento de ferramentas para a automatização do processo tradutório, além de sua evolução ao longo do tempo. Nas seções a seguir, são apresentados os conceitos de máquinas de tradução e outras aplicações acerca da sintetização da linguagem por máquina.

2.3.1. Máquinas de tradução: histórico, evolução e definições

No que concerne às tecnologias de PLN voltadas para o campo da linguagem, as máquinas de tradução (MT) estão entre as que mais se desenvolveram até então (AZIZ; CASTILHO; SPECIA, 2012; MELBY, 1997; SPECIA; FARZINDAR, 2010). Baseando-se na ideia de padronização da linguagem, as MTs utilizam a tradução automática, ou seja, um conjunto de instruções em uma máquina que realiza a tradução de conteúdo de uma língua para outra (AZIZ; CASTILHO; SPECIA, 2012). Neste sentido, as MTs entraram no mercado da tradução, como toda nova tecnologia, com a proposta de facilitar, diminuir o tempo de trabalho e aumentar demandas que se tornaram maiores, se tornando além da capacidade somente física de executar (FIEDERER; O'BRIEN, 2009; HUTCHINS, 2007).

Entretanto, os pressupostos de desenvolvimento de uma ferramenta que pudesse traduzir instantaneamente conteúdo de uma língua para a outra, eram (talvez ainda sejam) utópicos. Devido à complexidade computacional de realizar a tarefa de forma totalmente automatizada, ou seja, sem a interação de um humano para fazer a interpretação entre os idiomas, os *software*s desenvolvidos a princípio não atingiam nem mesmo um nível de resultado que permitisse uma breve compreensão do que se pretendia traduzir (CIVERA et al., 2004).

De modo geral, na década de 40, o conceito de recriar uma atividade humana, aparentemente simples em uma máquina, era o ponto de partida para os pioneiros do campo ainda inexplorado da linguística computacional (HUTCHINS, 2007). Grande parte do impulso para a criação de ferramentas que pudessem interpretar idiomas

instantaneamente se deu por conta do avanço militar e da necessidade de decifrar mensagens dos inimigos ao redor do mundo.

Partindo desse impulso inicial, o desenvolvimento em MTs se tornou uma vantagem na inteligência militar ao longo das décadas de 50 e 60, em que o conflito entre Estados Unidos e União Soviética na Guerra Fria se baseava, prioritariamente, na disputa tecnológica de cada país, principalmente em 1957 depois da disparada espacial da União Soviética com o lançamento da *Sputnik* e, consequentemente, iniciando a corrida tecnológica entre os países (CRACIUNESCU; SALAS; KEEFFE, 2004; HUTCHINS, 1986).

Sendo assim, o desenvolvimento das MTs era financiado pelas agências militares, pela sua utilidade em decifrar mensagens em outras línguas e também por garantir o pioneirismo tecnológico na área da computação (DORR; SNOVER; MADNAMI, 2010; SLOCUM, 1985).

Avançando para a década de 70, o desenvolvimento das MTs ganhou destaque no continente europeu por conta da unificação dos países da Europa para sanar a dificuldade multilíngue do continente. Tal realidade impulsionou o desenvolvimento de sistemas de tradução para, primeiramente, lidar com documentos administrativos, econômicos e técnicos (HUTCHINS, 1986). A MT foi o primeiro desenvolvimento não numérico aplicado à computação (HUTCHINS, 1986, 2007), fato que gerou contratempos e altas expectativas em meados da década de 60. Neste período, a qualidade da tradução foi medida pelo relatório ALPAC[14], estudo que resultou em uma demonstração infiel[15] da tradução automática, descontinuando drasticamente o desenvolvimento de MTs e, consequentemente, dos processadores de linguagem natural (FIEDERER; O'BRIEN, 2009; HUTCHINS, 1986, 2007).

Mesmo com o hiato no desenvolvimento das MTs, que ficou estagnado por quase vinte anos, a explosão da informatização nos anos 80 retomou o interesse das grandes

---

[14] Relatório ALPAC foi um experimento com MTs a respeito da precisão do conteúdo traduzido e foi um fracasso, descreditando a acurácia da tradução automática (FIEDERER; O'BRIEN, 2009; HALLIDAY, 2001).

[15] Fidelidade ou equivalência representa a qualidade da tradução tendo em base o conteúdo original (HALLIDAY, 2001; HUTCHINS, 2004).

empresas de tecnologia da informação[16] pela evolução da área (FIEDERER; O'BRIEN, 2009; HUTCHINS, 1986, 2007; SLOCUM, 1985). A continuação do desenvolvimento em ferramentas de linguagem por parte das grandes corporações não se deu apenas pelo *boom* da informatização, mas também por outros fatores linguísticos, teorias de investigação semântica através de *software*s (linguística de *corpus*) e implantação da Comunidade Econômica Europeia, que já havia se aventurado pelas MTs anteriormente (HUTCHINS, 1986, 2007; SLOCUM, 1985).

O advento da área da Linguística Computacional surgiu em meio a uma caracterização interdisciplinar, envolvendo Computação, Linguística e Estatística, a fim de desenvolver uma aplicação não numérica a ser compreendida pela máquina (OTHERO, 2006). Desta forma, são necessários recursos computacionais que superem os obstáculos da complexidade de aplicações como as MTs, como o desenvolvimento da Inteligência Artificial e o Aprendizado de Máquina. A busca por tornar a máquina mais "interpretativa", no campo semântico da linguagem, é um movimento necessário para que o processamento sintático funcione apenas como uma forma padronizada na criação de processos automáticos da língua no computador (BARRACHINA et al., 2009; CIVERA et al., 2004; KHUMMONGKOL; YOKOTA, 2015).

As MTs revolucionaram a história da tradução e dos tradutores, enfatizando a capacidade específica da profissão para a construção de máquinas mais efetivas no campo da linguagem. Assim, a reestruturação da área transformou e ainda transforma os procedimentos tradutórios, dando vazão a novos estilos de tradução baseados na tradução de máquina (HUTCHINS, 1986; SLOCUM, 1985).

2.4. Pós-edição em processamentos automáticos da linguagem

Esta subseção destaca a pós-edição das aplicações automáticas geradas para o PLN e servirá de embasamento para o capítulo de Metodologia, assim como no que diz respeito ao processo de pós-edição do *corpus* analisado.

Embora as aplicações de PLN tenham ganhado espaço ao longo do tempo como a solução necessária para sanar demandas volumosas dos profissionais de tradução e

[16] IBM com o experimento Georgetown em 1954, que traduziu quase cem frases do russo para o inglês instantaneamente (BUZATO, 2010).

sintetização de voz, somente o conteúdo automático não é suficiente para a compreensão final de certo conteúdo (SPECIA; FARZINDAR, 2010), não somente na tradução, mas na relação do PLN. Assim, o conteúdo inalterado gerado automaticamente pela máquina causa contratempos e é considerado adequado apenas para consultas rápidas a materiais (LIMA-LOPES; MORO, 2018 apud. LAURIAN, 1984).

Desta forma, as aplicações de PLN garantem a rapidez e a padronização dos conteúdos, tanto transcritos quanto traduzidos, interagindo com a memória, autocorreções da tradução automática, entre outros (AZIZ; CASTILHO; SPECIA, 2012; SPECIA; FARZINDAR, 2010). Com isso, a necessidade de interação do humano com a máquina se torna primordial, para definir as diretrizes em que o *software* vai se basear para a busca de termos e estabelecimento de parâmetros para autocorreção (ALLEN, 2001; AZIZ; CASTILHO; SPECIA, 2012; BARRACHINA et al., 2009).

Esta interação do humano com a máquina, no momento de edição de um conteúdo automático, leva o nome de pós-edição (PE), Esta etapa pode ser definida como uma revisão do conteúdo "bruto" gerado pela máquina, melhorando a precisão do produto final e também da própria máquina, se esta contar com recursos computacionais de aprendizado de máquina (ALLEN, 2001; LAURIAN, 1984).

O procedimento de PE em conteúdos gerados automaticamente se tornou comum após o massivo desenvolvimento neste campo, depois dos anos 80 (ALLEN, 2001), justamente pelo fato de contornar o problema com a precisão que o processo automático gera (FIEDERER; O'BRIEN, 2009). Embora tenha sido explorado em cada ferramenta, o processo de PE não segue uma metodologia específica de aplicação para todos os *software*s, pois os modelos propostos variam conforme a performance computacional e à interação do humano com a máquina (ALLEN, 2001).

Assim, mesmo com a não formalização de uma metodologia específica de PE, foram realizados alguns estudos a respeito de critérios primordiais para que o procedimento seja executado com o mínimo de apuração necessária dos conteúdos gerados automaticamente (LIMA-LOPES; MORO, 2018). Cada categoria de PE segue um critério específico, ou seja, uma espécie de objetivo final de cada processo, convergindo com as necessidades que cada conteúdo exige quanto no que diz respeito à precisão (LAURIAN, 1984), sendo 1) PE rápida, 2) PE mínima e 3) PE completa (LOFFLER-LAURIAN, 1996).

A PE rápida se caracteriza pelo critério de situação e do tipo do documento, ou seja, situação que engloba a verificação rápida de pontos críticos do processo automatizado de tradução e tipo do documento em relação à falta de necessidade de compreensão completa do conteúdo. Geralmente é utilizada para consultas rápidas a materiais que necessitem de poucas informações. Já a PE mínima revisa o conteúdo automático pontos de informações específicas ou o conteúdo como um todo, de forma reduzida se atentando a ideia principal. No caso da PE completa, o processo é executado com o critério de compreensão, isto é, o conteúdo é revisado com o objetivo de ser fiel ao original e compreendido minuciosamente (LOFFLER-LAURIAN, 1996).

Contudo, mesmo com a interação da máquina com o processo de PE, corre-se o risco de não obter-se um conteúdo de maneira compreensiva, uma vez que o PLN possui uma base sintática para o processamento, e não semântica. O fato da sintaxe ser a maneira da máquina "compreender" palavras é posto em prova para questões como: equivalência tradutória, diferentes fatores do que é considerada uma boa tradução em termos de sentido e compreensão e a própria compreensão de conteúdos processados pela máquina (HALLIDAY, 2001).

Assim posto, o próximo capítulo se dedica à metodologia de pesquisa adotada neste trabalho, buscando elaborar os modelos de treinamento para o software de PLN, Skylar, através de mediações da linguística computacional e da linguística de *corpus*.

CAPÍTULO 3 - METODOLOGIA

3. Objetivos e organização da metodologia

Este capítulo tem como objetivo principal apresentar os procedimentos metodológicos adotados para a elaboração dos modelos de treinamento linguístico do *software Skylar*. Em um primeiro momento, será apresentada a motivação e o contexto em que essa pesquisa se insere, enfatizando seus objetivos. Em um segundo momento, serão apresentadas as estruturas do *software* e suas funcionalidades com o objetivo de embasar o processo de levantamento e coleta dos dados utilizados na análise. A seguir, será exposto o procedimento metodológico de pós-edição do conteúdo gerado automaticamente pelo *software* por meio da transcrição, com o objetivo de se demonstrar o método em que o *corpus* foi coletado e configurado para a análise. Na sequência, com base no funcionamento do *software* e na configuração do *corpus* de pesquisa, serão apresentados os métodos de análise comparativa quanti-qualitativa, procurando estimar a capacidade de precisão do procedimento automático da aplicação.

A *Skylar* é uma ferramenta de tradução automático-simultânea que foi desenvolvida com inteligência artificial para realizar interpretação simultânea no formato de legendagem. Ela utiliza de algoritmos de *machine learning* para fazer a compilação dos dados através do reconhecimento de voz e o retorno desses dados traduzidos em formato de legendas durante a exibição ao vivo da videoaula ou de quaisquer outros conteúdos em formatos de vídeo que são submetidos a ela. É uma ferramenta autoral que foi desenvolvida especialmente para o instituto de EAD em sua estratégia de expansão internacional por meio de seus cursos, que só foi possível pela sua criação. As línguas disponíveis para utilização da *Skylar* são português, inglês e espanhol (em quaisquer configurações de partidas e chegadas).

A ferramenta reconhece a voz no idioma de partida, transforma a informação do áudio em texto no formato de transcrição. O algoritmo, por sua vez, faz a temporização da legenda e a tradução da transcrição simultaneamente, fornecendo o resultado da legenda traduzida ao usuário com poucos segundos de espera. Além deste procedimento, a IA contida no software precisa ser treinada por especialistas humanos, a fim de mitigar possíveis erros dentro do reconhecimento de voz. Para tanto, foi elaborada uma metodologia de apuração do conteúdo processado pela *Skylar* a fim de se propor

treinamentos específicos explorando o que a ferramenta mais tem dificuldade em processar, causando erros e inconsistências nas legendas.

Tais inconsistências geram contratempos no entendimento das legendas por parte dos usuários, causando um problema na precisão do conteúdo processado. Por precisão, entende-se a relação entre o resultado gerado automaticamente pela máquina em contraste com o resultado obtido depois de um tratamento de PE que será exposto nesta seção. Para tanto, serão apresentadas as métricas utilizadas a nível quantitativo e qualitativo nesta seção.

Para fins quantitativos, serão utilizadas duas equações como métricas, a saber: a WER – *Word Error Rate* (KLAKOW; PETERS, 2002), para contabilização do índice de erros em um comparativo entre o conteúdo original e o ideal; e a LD – *Levensthein Distance* (LEVENSTHEIN, 1966), para medir a distância entre os caracteres dentro do comparativo. Utilizando como base as porcentagens do WER e as estatísticas de LD, os tipos de alterações encontradas serão classificados com o objetivo de levantar quais são as porcentagens dos erros apresentados pela máquina em relação ao conteúdo ideal. Além disso, destacará os erros mais frequentes, classificando-os e setorizando-os entre momentos da aula transcrita. Desta forma, será possível identificar a partir de quais categorias os modelos de treinamento seriam mais efetivos para a *Skylar*.

Por meio das métricas, apresentadas, será exposta, após a identificação e apuração dos dados, uma análise qualitativa de termos mais frequentes e seus contextos, de forma a traçar um movimento possivelmente padronizado no discurso do locutor, no caso, o professor. Esta análise se dará com o uso da ferramenta de análise de *corpus Sketch Engine*[17].

Embora o *software* se apresente como um processador de legendas automático em diferentes línguas, a língua partida usada para análise de precisão, nesta pesquisa, será o português. Ou seja, os modelos serão gerados a partir dessa transcrição. Todas as análises aplicadas para verificação de precisão serão expostas pela transcrição, pois, por regra do *software*, uma vez que a língua fonte é captada e transcrita, o mesmo conteúdo é traduzido automaticamente para as outras línguas disponíveis, ou seja, as correções da transcrição

[17] https://www.sketchengine.eu/

são espelhadas na tradução. Esta é uma funcionalidade específica do *software* e é onde surgem mais equívocos, pois os sistemas linguísticos de partida e chegada são completamente diferentes, colocando em questão se a precisão da língua de partida realmente deixa a língua de chegada mais precisa[18].

[18] Busca-se testar neste estudo se é realmente efetivo um modelo de treinamento que padroniza todo o conteúdo do que é transcrito e replica para a tradução. As análises de contexto serão inseridas em trabalhos futuros para se obter mais precisão nessa replicação de correção.

3.1. Compilação do *corpus* e *software* de análise

O *corpus* elaborado para esta pesquisa foi gerado a partir da ferramenta de transcrição, legendagem e tradução simultânea *Skylar*, uma ferramenta autoral[19] que está sendo desenvolvida para o instituto de EAD Pecege[20], com o intuito de proporcionar legendas simultâneas e traduzidas em aulas ao vivo. O primeiro objetivo do *software*, motivo principal de sua criação, era gerar legendas e enviá-las para a transmissão ao vivo destas aulas. Porém, além dos percalços tecnológicos que a criação de uma ferramenta dessa complexidade ocasionou, houve uma questão ainda maior que acarretou a motivação do treinamento linguístico do *software*.

A ferramenta foi desenvolvida com foco em tornar-se um sistema especialista de reconhecimento de voz, transcrição e tradução simultânea utilizando IA em seu processamento. Conforme será descrito nas próximas seções, a ferramenta reconhece a voz, transcreve, temporiza e segmenta no formato de legendas, e traduz simultaneamente o conteúdo nela processado.

Uma vez que se busca trabalhar com ferramentas de reconhecimento de voz, o conteúdo está sujeito aos mais variados tipos de erros, como erros de reconhecimento fonético (Ex.: Automático: eu cansei > Ideal: eu tracei), ausência de sílabas ou palavras (Ex.: Automático: colchas em relação aos dados > Ideal: Como eu falei para vocês, minha desconfiança), nomes próprios não reconhecidos (Ex.: Automático: bebê já tem uma pesquisa > Ideal: o IBGE já tem uma pesquisa), substituição de palavras que mudam o sentido, etc., fatores que serão discorridos na seção de análise dos resultados. Por essa razão, a *Skylar* foi desenvolvida com mecanismos de IA para que estes erros fossem reconhecidos e tratados, autocorrigindo quaisquer que fossem as ocorrências equivocadas provenientes da falha reconhecimento de voz.

---

[19] A *Skylar* (www.*Skylar*.ai) foi idealizada e desenvolvida pela autora, MSc Lucas Guerreiro e Prof. Dr. Daniel Y. Sonoda, com aporte financeiro do Pecege. Doravante será exposta sua estrutura e seu funcionamento.

[20] O Pecege (www.pecege.com) é uma associação que engloba iniciativas de educação, pesquisa e tecnologia. Entre suas marcas, está a *Skylar* (www.*skylar*.ai), *software* para legendagem e tradução simultânea, e o MBA USP/ESALQ (www.mbauspesalq.com), cursos de MBA a distância na área de gestão, sendo uma de suas aulas utilizada nesta pesquisa como será doravante exposto.

Desta forma, surgiu a motivação para que houvesse uma forma mais eficiente de se avaliar os erros gerados pela ferramenta e, consequentemente, a inserção dessa maneira no algoritmo de autocorreção do *software*, ou seja, ocasionando em uma metodologia específica para a elaboração de modelos para o algoritmo de *machine learning*[21] aplicado a *Skylar.* Esta metodologia busca a apuração de dados linguísticos-estatísticos acerca do conteúdo gerado automaticamente pelo *software*, viabilizando assim a possibilidade de desenvolvimento de modelos linguísticos que auxiliariam no processo automático da passagem do áudio para texto, treinando a IA composta no *software,* para que haja o processo de identificação dos erros e autocorreção dos conteúdos considerados equivocados.

Como exposto, o processo automático da transcrição mediado por um *software* especialista, como no caso da *Skylar*, necessita de uma análise linguística eficiente para o levantamento dos erros. Por esta razão, a primeira parte da metodologia refere-se ao ato da correção manual do conteúdo gerado automaticamente (pós-edição), elaborando um formato ideal de como este segmento deveria ser. Esta correção é feita por analistas humanos especializados[22]. Como será demonstrado nas demais seções, o *corpus* está no formato de legendas considerando apenas a transcrição do conteúdo.

Nas próximas seções, serão abordados os conteúdos de estrutura e funcionamento da *Skylar*, processos metodológicos de pós-edição e os processos metodológicos das análises linguísticas-estatísticas desenvolvidas para esta pesquisa. Esses processos descritos utilizam equações da linguística computacional, na parte quantitativa, e parâmetros da LC para a parte qualitativa.

---

[21] *Machine learning* é uma técnica computacional desenvolvida para aprender tarefas e dar as respostas corretas para uma questão, independente do seu contexto (GOLDBERG, D.E.; HOLLAND, 1988, p. 95).
[22] Estes especialistas são parte do time de Inteligência Artificial do Pecege/*Skylar*. Nesta pesquisa as correções foram elaboradas pela autora.

3.2. Características estruturais da *Skylar*

Para entender o processo de legendagem executado pela *Skylar*, devemos partir da definição de legendagem. Essencialmente, a legendagem é definida "como prática de tradução que consiste na apresentação de um texto escrito, geralmente na parte inferior da tela, que busca recontar o diálogo original dos falantes" (CINTAS; REMAEL, 2007, p. 8), na intenção de promover conteúdo audiovisual acessível para não proficientes da língua original ou que possuam algum tipo de deficiência auditiva. Para tal, devem-se seguir alguns critérios no que concerne a abordagem linguística:

> "- Uso de unidades sintáticas simples e livre de repetições (texto condensado, reduzido e simplificado) e em norma culta, sem erros gramaticais, uma vez que "as legendas servem de modelo para o letramento";
>
> - Segmentação e distribuição linear e em blocos com base em unidades gramaticais e de sentido;
>
> - Informações relevantes escritas em imagens devem ser traduzidas em caixa alta, e músicas, quando de conteúdo semanticamente relevante, devem ser traduzidas em itálico;
>
> - Tempo de exibição das legendas na tela: deve obedecer à velocidade média de leitura, com duração mínima de 1 segundo e máxima de 7 segundos;
>
> - Número máximo de linhas: deve ser limitado a duas e, preferencialmente, a linha de cima deve ser mais curta que a linha de baixo, para facilitar a visualização da imagem e reduzir o movimento dos olhos" (SPOLIDORIO, *apud* NAVES *et al*., 2016 p.42).

Tais definições foram estabelecidas como procedimentos para a elaboração de legendas pelos profissionais da tradução audiovisual, de forma a funcionar como um manual para que as legendas sejam executadas de forma que as pessoas consigam acompanhar o conteúdo audiovisual apresentado, da mesma forma que aquele que entende o que ouve (CINTAS; REMAEL, 2007, p. 8–9). Para o foco do desenvolvimento da *Skylar*, tendo como base as abordagens linguísticas da legendagem, dois itens dos cinco apresentados anteriormente são os pilares estruturais do *software*: (i) a compilação de unidades sintáticas simples e livres de repetições e (ii) a distribuição e segmentação lineares baseadas no sentido.

Esses dois itens, a compilação de unidades sintáticas simples e livres de repetição e segmentação lineares baseadas no sentido, são itens centrais para a estruturação do processamento de linguagem executado pela *Skylar*. O primeiro porque a fala do locutor não é livre de repetições, logo, o *software* necessita tratar o que é processado para que o segmento não fique com repetições provenientes de coloquialismos da fala. O segundo, porque os segmentos não são executados em um mesmo local, ou seja, a mesma linearidade do que está sendo dito é segmentada, fazendo com que o *software* não conecte as duas frases, oferecendo sentidos mais exatos. Portanto, os modelos de treinamento devem recair de forma mais reforçada sobre esses dois pontos.

Nessa perspectiva, um dos maiores desafios da automatização das legendas está em seguir estes critérios linguísticos-audiovisuais para que o conteúdo da leitura seja compreensível por quem assiste as videoaulas, pois as legendas também "servem de modelo para o letramento" (NAVES et al., 2016, p. 42). A legenda como modelo de letramento significa que, através delas, pode-se obter um modelo para letramento no sentido do tipo de informação ou conhecimento que está sendo passado através desta ponte. Neste caso, a legenda também faz parte do aprendizado do aluno e do seu entendimento tanto quanto o conteúdo em si.

Um dos propósitos da elaboração de modelos linguísticos para legendagem propostos por esta pesquisa é gerar um conteúdo de legenda próximo ao realizado por um humano (transcrição ideal), apresentando modelos efetivos para o treinamento da IA. Buscando seguir as recomendações da geração de legendas, a *Skylar* executa uma sequência de quatro procedimentos básicos, conforme consta na Figura 2. Dentro desses procedimentos, principalmente no momento de transcrição e segmentação das legendas, é que se encontram as regras descritas anteriormente.

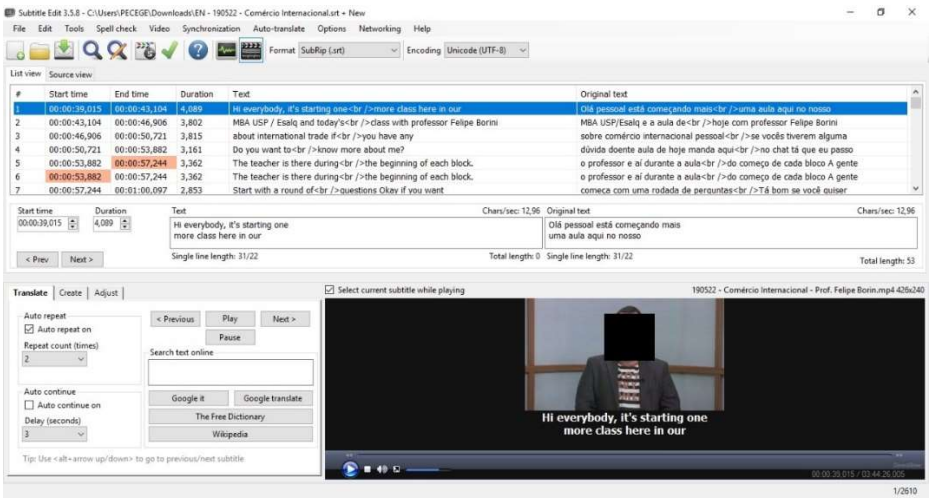
Figura 2 - Procedimentos básicos de execução da *Skylar*

Fonte: Fluxo da Skylar elaborada pela autora.

Esta sequência representada na figura 2 demonstra como o *software* executa as legendas automaticamente, de forma a entregar o conteúdo para o usuário. O *software* realiza o processamento automático através do reconhecimento de voz, gerando uma transcrição. Além disso, segmenta as legendas e insere o tempo, gerando uma tradução automática através da transcrição evidenciada no passo 2. O processo de legendagem apresentado pela *Skylar* mostra-se inovador, uma vez que, atualmente, este processo é feito de forma integralmente manual, na qual o legendador executa a transcrição (ouvindo o áudio e transcrevendo à mão), a tradução (com auxílio de CAT Tools[23]), a segmentação dos blocos de legenda – respeitando as normas linguísticas e audiovisuais, além da formatação de tempo (parando a cada milissegundo para ajustes manuais). Essas etapas são executadas, geralmente, com uma junção de *software*s que auxiliam no processo, resultando na revisão final, por meio de ferramentas como o *Subtitle Edit*[24] (figura 3).

[23] *CAT Tools* (Ferramentas de tradução assistida por computador) são ferramentas de auxílio à tradução que possuem uma série de funcionalidades para deixar o processo de tradução mais eficiente, como glossários, memórias de tradução, tradução automática, etc. (MUNDAY, 2012).
[24] *Subtitle Edit* é um editor de legendas gratuito e de código aberto para criar, editar, ajustar ou sincronizar legendas para vídeos. Disponível em: https://subtitle-edit.br.uptodown.com/windows

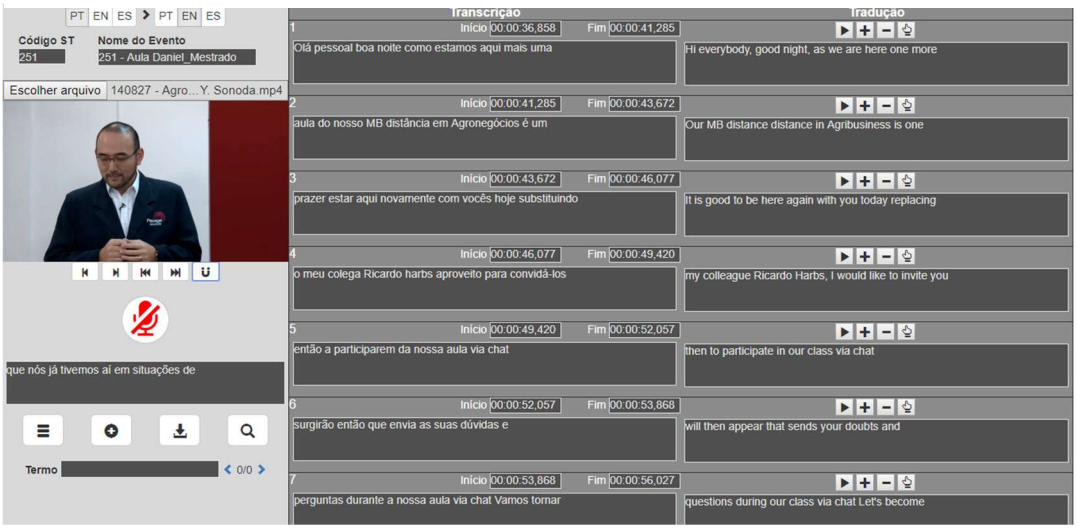
Figura 3 - Exemplo de software para auxílio de legendagem – *Subtitle Edit*



Fonte: Captura de tela da aplicação *Subtitle Edit*.

A fim de comparação e demonstração, a interface (figura 5) da *Skylar*, na qual os procedimentos da figura 2 ocorrem, caracteriza a dinamicidade do *software*, ou seja, demonstra quantas opções são expostas ao usuário enquanto uma legenda é gerada automaticamente, diferente do exemplo anterior em que o profissional necessita executar todas as funções de forma manual. Na figura 4, por sua vez, podem ser observadas as posições em que os segmentos se encontram e os quadros de edição que são utilizados pelos usuários no momento da pós-edição. Já a figura 5 expõe as funcionalidades por setores.

Figura 4 - Geração de conteúdo pela Skylar



Fonte: Captura de tela da aplicação *Skylar*.

Figura 5 - Setores da interface da Skylar destacadas

Fonte: Captura de tela da aplicação *Skylar*.

A Figura 5 expõe os setores destacados do *software*, pela ordem de funcionamento para a geração de conteúdo das legendas, conforme descrito no quadro 1.

Quadro 1 - Descrição das funcionalidades do software

Área	Descrição
A	*Upload* do arquivo de vídeo para gerar o conteúdo.
B	Escolha do idioma: no primeiro grupo o idioma que será transcrito (idioma de partida) e no segundo bloco o idioma que será traduzido (idioma de chegada).
C	Ativação do microfone para o reconhecimento de voz, automaticamente reproduz o vídeo.
D	Quadro para geração da transcrição.
E	Botões com funcionalidades adicionais: voltar para o menu, novo, *download* e busca por legendas no banco de dados.
F	Pesquisa de termo específico, no qual se busca o termo, contabiliza a quantidade de termos transcritos e se é possível navegar entre eles, demonstrando o segmento correspondente ao lado.
G	Coluna da transcrição, na qual aparecem os conteúdos gerados depois de segmentados na parte D.
H	Indicação do segmento e do tempo correspondente de legenda. O corte de segmento acontece pós a transcrição realizada na parte D.
I	Coluna da tradução, na qual estão os conteúdos gerados depois de transcritos na parte D e separados na parte H.
J	Botões com funcionalidades adicionais: ir para segmento, inserir segmento abaixo, remover segmento e juntar com segmento acima.
K	Quadros de transcrição e tradução nos quais são executadas as edições das legendas automáticas.
L	Botões com funcionalidades adicionais: voltar 10s, avançar 10s, adiantar legendas, atrasar legendas e *auto-scroll*.

Fonte: Elaboração da autora.

No Quadro 1 podem ser observadas as funcionalidades do *software*, tendo suas funções descritas na coluna “Descrição”. Foram expostas na ordem de execução,

demonstrando o processo de geração de legendas em sua totalidade[25]. Nas funções A e B, o usuário escolhe o vídeo que deseja legendar e escolhe os idiomas de partida e chegada[26], respectivamente, sendo a primeira interação necessária para geração de conteúdo no *software*. Na C, há o ícone de microfone, no qual o usuário deve ativar para iniciar o processo de legendagem do vídeo escolhido. Nesse momento, o vídeo é automaticamente ativado, para que haja a sincronia nas legendas a serem geradas[27]. Na caixa D, a transcrição é gerada antes de ser segmentada nas caixas da coluna G. Conforme o *software* processa o áudio, esta caixa é preenchida pelo conteúdo transcrito. Nas áreas E e F, há algumas funcionalidades operacionais do *software*, sendo a segunda uma aplicação voltada diretamente à pesquisa de termos da transcrição. Conforme os termos vão sendo pesquisados e encontrados, o vídeo se auto redireciona para o momento em que aquele termo foi falado. Através deste recurso é possível contabilizar a frequência de termos recorrentes, fator determinante para o cálculo de colocados que será realizado na análise qualitativa deste trabalho, além de dar margem para a criação de uma transcrição interativa[28].

A coluna G, indicada na Figura 3, é a funcionalidade que garante os segmentos de transcrições segmentados, além de ser a área para realizar-se a pós-edição do conteúdo gerado. Na H, constam as informações de temporização e segmentação das legendas que são geradas automaticamente. Nesta seção ocorre a edição de legendagem propriamente dita, ou seja, a correção de temporização e segmentação (com os botões de adição/exclusão de legendas, junção de legendas da J). Na coluna I, há o mesmo conteúdo apresentado na transcrição, mas já contendo a tradução na língua determinada anteriormente. Na K, espaços onde são executadas as edições correspondentes para os

[25] O processo de geração de legendas automáticas está sendo descrito a partir da técnica para vídeos gravados. Porém, o *software* também pode ser executado com vídeos ao vivo e com captação de som externo (apenas transcrição).
[26] Os idiomas de partida e chegada podem variar entre português, inglês e espanhol, também tendo a possibilidade de ser o mesmo idioma de partida e chegada, no caso de legenda apenas transcrita do áudio.
[27] Nesse caso, por estarmos lidando com vídeos gravados, o som é obtido através de um microfone virtual instalado no computador que altera as configurações de som do computador e transforma a captação de som do computador através do *software*.
[28] Funcionalidade futura no desenvolvimento do *software* a fim de o tornar mais efetivo em pesquisas de LC com transcrições de áudio.

possíveis erros que podem acontecer durante o processo automático de transcrição. Na última seção, L, há alguns botões com funcionalidade adicionais para o *player*.

Isto posto, as duas próximas seções se dedicam à apresentação do processo de pós-edição do material gerado pela *Skylar,* a fim de que o *corpus* fosse preparado para as análises posteriores.

3.3. Processo de pós-edição (PE) da transcrição automática

Embora o *software* utilizado realize a geração de legendas de forma automática, algumas métricas deveriam ser estabelecidas para que fosse possível observar e identificar a fidelidade do conteúdo transcrito gerado pela ferramenta, obtendo-se resultados relevantes para a criação de modelos para o treinamento do *software*. A questão de fidelidade ao conteúdo é algo relevante para a área da tradução. Abordar este tópico significa dizer que um texto traduzido deve ser o máximo fiel ao seu texto de origem, depois de ter sido transposto para outro idioma (BRITO, 2010). Nesta pesquisa especificamente, considera-se apenas a fidelidade dentro da transcrição gerada pelo software em relação ao original – captado através do reconhecimento de voz. Porém, essa tem uma grande relação com a tradução, uma vez que o *software* "espelha" a transcrição para mobilizar um processo de tradução automática.

Leva-se em consideração, na realização da pós-edição (PE) de um conteúdo automático, a relação do contexto e do co-texto. O contexto aqui explorado segue a definição da inter-relação entre as palavras e a informação principal contida na sentença, isto é, a representação semântica que o conteúdo oferece em ambas as línguas (HALLIDAY, 2001). A relação do co-texto se insere na união entre o texto corrido transcrito e a legenda, sendo que ambos trabalham de formas conjuntas para a realização do conteúdo final para o usuário.

Conforme exposto anteriormente, todo processo automático de linguagem natural pode ser gerado com erros na transcrição. Por isso, a elaboração dos modelos linguísticos de treinamento do *software* busca efetivar a precisão necessária para que o conteúdo seja veiculado aos que o consomem e seja tão efetivo quanto dos nativos que estão absorvendo o mesmo conhecimento. Evidentemente, a tradução não se define apenas nessa troca do código linguístico e sim uma transposição cultural e linguística de certo conteúdo, o qual tem a intenção de ser idealmente fiel ao seu original (HALLIDAY, 2001). Porém, dentro do processamento do software em questão, a relação de fidelidade do conteúdo original para com sua origem (o reconhecimento da voz) e a transposição para a tradução automática, se torna primordial ao menos no nível da transformação linguística, não sendo considerados, nesse momento, outros aspectos apresentados pela técnica de tradução.

A iniciativa de se criar um processo metodológico para a PE teve como ponto de partida o artigo apresentado por Allen (2001), que expõe resultados positivos na aplicação de uma metodologia em um *software* de PLN. Allen (2001) apresenta em seu trabalho uma redefinição dos propósitos acerca da tradução automática dentro do processo natural frequentemente utilizado, utilizando uma metodologia própria de PE em um *software* de tradução elaborado com algoritmos de *machine learning.*

O mesmo princípio é replicado, neste ponto, na intenção de conceder uma melhoria na qualidade e velocidade nos processos tradutórios para os profissionais que manipulam o *software* e propõem adequações do processo automático, acompanhamento da maior demanda de conteúdos traduzidos e o levantamento de pontos de melhorias do *software* indicados em cada etapa. Sendo assim, o desenvolvimento de um processo metodológico de PE (ALLEN, 2001) para a ferramenta foi desenvolvido para compreender o processo automático, visando o resultado mais preciso possível no PLN da ferramenta (LIMA-LOPES; MORO, 2018, p. 58). Neste sentido, a precisão na transcrição gerada é o objetivo central das análises que serão apresentadas. A precisão, ou seja, a relação entre o resultado gerado automaticamente pela máquina, em contraste com o resultado obtido depois de um tratamento de PE, será exposto nesta próxima seção.

3.3.1. Estabelecendo processos metodológicos de PE

Processos metodológicos a respeito da PE do conteúdo automático devem ser estabelecidos antes do processo de análise com métricas e parâmetros, pois, ao identificar tais erros[29], é possível inferir os objetos que devem ser priorizados dentro de um modelo linguístico de treinamento para o *software* (ALLEN, 2001), retornando um conteúdo mais fidedigno ao aluno submetido às legendas durante a aula. Conforme citado anteriormente, sistemas de reconhecimento de voz podem gerar erros ao reconhecer e transcrever o que está sendo dito, deixando o conteúdo inteiramente automático não fiel ao que foi realmente falado. Nesse caso, o retorno de um conteúdo correto ao aluno se faz necessário visto que este absorverá o conteúdo da aula através das legendas propostas.

[29] Este processo de divisão dos tipos de erros será apresentado doravante na análise quantitativa dos dados.

Nessa perspectiva, os elementos metodológicos de PE propostos são responsáveis por apontar os problemas ocorridos durante o processo automático da geração da transcrição, sendo possível determinar os tipos de erros existentes no conteúdo e em que ponto os modelos devem atuar para melhorar a interação do discurso do locutor com a interpretação do interlocutor. Esta interação entre estes objetos é um dos pontos centrais para os resultados a serem expostos, pois é o que garante que o interlocutor (o aluno) compreenda o conteúdo passado pelo locutor (o professor) da mesma forma que o nativo que não necessita do suporte de legendas[30].

Seguindo os critérios de execução do *software* (Quadro 1), foi estabelecido o processo de PE adequado, ou seja, que geraria melhor contribuição para a formação dos modelos linguísticos, que segue a sequência estabelecida no Quadro 2.

Quadro 2 - Etapas metodológicas de PE propostas para a aplicação Skylar

Etapa	Descrição
1. Transcrição Assistida	Entrada do conteúdo por reconhecimento de voz e análise do procedimento durante essa captura. Identificação de segmentos críticos na transcrição e tradução através de uma revisão simultânea à coleta do conteúdo.
2. Revisão e PE	Edição dos pontos críticos da etapa 1 e PE completa do conteúdo.
3. Finalização	Revisão completa do conteúdo transcrito e traduzido[31].

Fonte: LIMA-LOPES; MORO, 2018, p. 60.

As etapas metodológicas de PE demonstradas no quadro 2 seguem, portanto, uma sequência de procedimentos a serem executados na *Skylar* para realizar o acompanhamento do processamento do *software*, a análise do que foi gerado e a correção das legendas. Através do método de PE, é possível identificar os erros da transcrição automática e chegar a um segmento ideal.

[30] Do ponto de vista da elaboração do software e das premissas de PE, a tradução é explorada aqui como a transformação de um código no outro, não levando em consideração o aspecto cultural dos idiomas. A ideia futura de desenvolvimento da IA é que ela perceba não somente os aspectos linguísticos padronizados e sim as interpretações mais profundas dos idiomas para que possa executar as inversões de forma correta.

[31] Este quadro foi elaborado na publicação original em que se considerava o processo de tradução do conteúdo. Para esta pesquisa, serão considerados apenas os processos de transcrição.

Embora o processamento de fala não se baseie nas mesmas premissas que os processos escritos – considerando a intenção de mecanização da fala por alguma sequência mediada por um *software*, o processo exato da aplicação e a avaliação de modelos linguísticos se desenvolve como uma referência para que a execução desta mecanização se aprimore com a maior relação de precisão possível, propiciando a melhor experiência ao usuário que irá consumir esta leitura do conteúdo oralizado. O conteúdo falado pelo locutor pode não seguir uma linearidade, ou até mesmo conter hesitações na fala e outras ocorrências que serão discutidas posteriormente, por isso a PE se torna necessária no processo anterior à análise dos segmentos em busca da precisão.

Um ponto contrário e singular desta metodologia em relação à de Allen (2001), que apresenta a análise em um *software* não especializado, é o de que as correções efetuadas na *Skylar* pelos analistas são armazenadas em processos de IA, fornecendo o auto aprendizado e autocorreção para os conteúdos gerados. Uma forma desenvolvida para "prever" o que o locutor pode expor, se baseando na linha de pensamento já executado previamente por outros exemplos[32].

Desta forma, as seções 3.4 e 3.5, a seguir, expõem as análises quantitativas e qualitativas a serem realizadas com os dados gerados, respeitando o processo metodológico apresentado nesta seção (Quadro 2) e as etapas de análise desenvolvidas. Novamente, levando em consideração o processo da transcrição na língua fonte para estabelecer as análises e modelos de treinamento.

[32] Em um *software* com quaisquer abordagens de IA (técnica de *machine learning),* os exemplos treinados pelos algoritmos são armazenados e replicados novamente quando a mesma estrutura textual aparecer (neste caso tratando apenas de *software* de PLN, mas pode ser aplicado em diversos fins); ou outra estrutura que ainda não lhe foi apresentada, onde busca encaixar em um padrão. Desta forma, o *software* "aprende" com as estruturas inseridas no sistema, podendo ser dentro do mesmo contexto ou não. Esta relação será exposta com mais detalhes posteriormente.

3.4. Procedimentos de Análise quantitativa de transcrição automática

A abordagem metodológica quanti-qualitativa utilizada neste estudo foi dividida em três fases, sendo 1) análise quantitativa, 2) análise qualitativa e 3) aplicação dos dados estatísticos. Nesta última, juntamente com os dados qualitativos, para a elaboração de modelos linguísticos de treinamento, gerando a análise dos conteúdos.

No que tange à 1) análise quantitativa, serão extraídos dados mensuráveis e específicos do conteúdo gerado automaticamente no formato de legendas e serão analisadas a) a produção do conteúdo ideal em relação ao que foi gerado automaticamente, b) a aplicação da métrica WER para cálculo de diferença entre as palavras do automático com as palavras do ideal, c) a aplicação da fórmula de *Levensthein* (1966) (LD) para comparação de quantidade de caracteres distintos entre o que foi gerado automaticamente e o ideal; e, por fim, o d) levantamento dos segmentos que possuem as melhores (WER menor ou igual a 49%) e piores (WER maior que 49%) pontuações através dos cálculos de WER, buscando obter os fatores que serão considerados na análise qualitativa.

3.4.1. Procedimento a - A produção do conteúdo ideal em relação ao que foi gerado automaticamente

Em um primeiro momento, foram coletados os conteúdos gerados automaticamente pela *Skylar* e depositados em uma tabela conforme exposto no Quadro 3.

Quadro 3 - Disposição do conteúdo, utilização da metodologia de PE para geração e correção dos dados (relação entre transcrição automática e transcrição ideal)

N. Segmento	Transcrição automática	Transcrição ideal
1	Olá pessoal boa noite como estamos aqui mais uma	Olá, pessoal, boa noite. Começamos aqui mais uma
2	aula do nosso MB distância em Agronegócios é um	aula do nosso MBA a distância em Agronegócios.
3	prazer estar aqui novamente com vocês hoje substituindo	É um prazer estar aqui novamente com vocês hoje, substituindo
4	o meu colega Ricardo harbs aproveito para convidá-los	o meu colega Ricardo Harbs. Aproveito para convidá-los

5	então a participarem da nossa aula via chat	então a participarem da nossa aula via chat
6	surgirão então que envia as suas dúvidas e	sugiro então que envie as suas dúvidas e
7	perguntas durante a nossa aula via chat Vamos tornar	perguntas durante a nossa aula via chat. Vamos tornar
8	nossa aula mais dinâmica e participativo um recado	nossa aula mais dinâmica e participativa. Um recado

Fonte: Elaboração da autora.

Nota-se que o conteúdo gerado automaticamente possui alguns erros em relação ao ideal. Desta forma, foi disposto no Quadro 3 o segmento correspondente à sequência (N. do segmento), a coluna do conteúdo automático (Transcrição automática) e a coluna para a transcrição ideal, ou seja, a transcrição corrigida de acordo com o que foi falado pelo locutor.

A produção da coluna ideal foi realizada de forma manual, por meio de processos tradicionais de transcrição. O processo de transcrição e PE do conteúdo seguiu a metodologia de PE citada na seção de pós-edição.

3.4.2. Procedimento b - A aplicação da métrica WER – *Word Error Rate*

WER – *Word Error Rate* é uma métrica para analisar a sequência de palavras e é utilizada para medir a performance em sistemas de reconhecimento de voz e máquinas de tradução (KLAKOW; PETERS, 2002). O objetivo principal da aplicação desta equação (Eq. I) ao conteúdo gerado é o de comparar trechos em termos de comprimento de palavras e analisar se a sequência de palavras se alterou no sentido quantitativo. Assim, apurar em termos exatos, quais foram as diferenças entre o conteúdo gerado automaticamente e o alterado. Desta forma, foi possível verificar o quão preciso o *software* se mostrou, se baseando no comprimento das palavras, conforme aplicação exposta no Quadro 4.

$$WER = (S + D + I) / N \qquad (I)$$

S = Substitutions
D = Deletions
I = Insertions
N = Number of words

Quadro 4 - Aplicação da equação WER

N. Segmento	Transcrição Automática	Transcrição ideal	WER
1	Olá pessoal boa noite como estamos aqui mais uma	Olá, pessoal, boa noite. Começamos aqui mais uma	S = 1 I = 1 N = 8 WER = (1+0+1) / 9 = **0,22** **Errou 22%**

Fonte: Elaboração da autora

Como pode ser visto na Eq. I, a fórmula para o cálculo da WER é composta por uma sequência matemática na qual se leva em conta, em relação ao conteúdo original, o número de substituições (S, *substitutions*), o número de exclusões (D, *deletions*), inserções (I, *insertions*) e o número de palavras (N, *number of words*). A fim de medir a performance de aplicações com reconhecimento de voz (*ASR – Automatic Speech Recognition applications*), a fórmula WER irá calcular a diferença entre o que foi gerado automaticamente com a transcrição ideal, funcionando como uma referência para a geração de conteúdo automático.

No Quadro 4, há um exemplo de como a equação WER (Eq. I) será aplicada nesta pesquisa e, pode-se observar, que ao executar a equação, temos como resultado uma porcentagem relacionada à quantidade de erros que o processo automático produziu: 22% e, sendo a porcentagem de acertos, 78%. Com a análise total do conteúdo, chegaremos aos números de acurácia que o processamento automático consegue e não consegue atingir.

Para que as análises com a métrica WER fossem executadas, foi elaborado um algoritmo[33] que executa as análises de forma automática nos segmentos recortados para esta pesquisa.

3.4.3. Procedimento c - A aplicação da *Levensthein Distance*

Levensthein Distance é utilizada para medir a similaridade entre duas sequências de caracteres, a fim de encontrar a proximidade entre elas (LEVENSTHEIN, 1966). A aplicação desta métrica é utilizada, prioritariamente, para o encontro de dois caracteres semelhantes com o objetivo de torná-los mais próximos, ou seja, encurtando a distância entre eles. A Figura 6 exemplifica a aplicação de LD para o cálculo de distâncias entre as palavras *kitten vs. sitting* e *Saturday vs. Sunday.* Os caracteres se diferem e se assemelham em alguns pontos do exemplo, sendo atribuídas distâncias conforme ocorrem as semelhanças e diferenças.

Figura 6 - Exemplos de aplicação de LD

		k	i	t	t	e	n
	0	1	2	3	4	5	6
s	1	1	2	3	4	5	6
i	2	2	1	2	3	4	5
t	3	3	2	1	2	3	4
t	4	4	3	2	1	2	3
i	5	5	4	3	2	2	3
n	6	6	5	4	3	3	2
g	7	7	6	5	4	4	3

		S	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

Fonte: PUTRA; SUPRIANA, 2015

A Figura 6, acima, exemplifica duas matrizes que comparam as palavras *kitten* com *sitting* e *Saturday* com *Sunday* demonstrando como a pontuação de LD é executada na comparação de palavras próximas em sua ortografia e fonética. Nesse caso, a LD é contabilizada utilizando-se a quantidade de letras total que as palavras possuem e atribuindo o mesmo número, a depender da posição em que elas se encontram. Esta forma

[33] Elaborado em colaboração por MsC Lucas Guerreiro, Mestrado em Ciências da Computação pela Universidade Estadual Paulista – UNESP.

de cálculo de LD é utilizada para a distância entre as *strings*[34], em um programa de computador.

Neste trabalho, porém, será aplicada de uma maneira modificada para determinar o encontro de caracteres (letras) que são semelhantes em seus sufixos e prefixos, tendo como princípio básico a LD, de forma a determinar se a porcentagem de erro determinada pela aplicação da WER está relacionada com a proximidade dos sons entre as palavras da sequência, conforme mostra o Quadro 5.

[34] Variável utilizada em linguagem de programação para representação de caracteres.

Quadro 5 - Aplicação das abordagens WER e LD

N. Segmento	Transcrição Automática	Transcrição ideal	WER	LD
1	Olá pessoal boa noite com**o** **est**amos aqui mais uma	Olá, pessoal, boa noite. Começamos aqui mais uma	S = 1 I = 1 N = 8 WER = (1+0+1) / 9 = **0,22** **Errou 22%**	**o st = 4** (com espaço) Olá**,** pessoal, boa noite**.** = **3 (pontuações)** LD Total = 7

Fonte: Elaboração da autora.

Como mostra o Quadro 5, entre as duas sequências apresentadas, houve uma diferença de 4 caracteres entre a transcrição automática e a ideal, considerando que se manteve o "e" da frase original e foram deletados o "o", "s" e "t", considerando o espaço entre eles. Além disso, a pontuação também é fator determinante para o cálculo de distâncias entre as frases, sendo incluídos 3 caracteres de pontuação entre as palavras "Olá", "pessoal" e "boa noite", além do "." no final da frase. Em relação ao WER, a transcrição automática apresentou apenas 22% de erro em relação a ideal e, através da LD, pode ser observado que a distância total foi de 7 caracteres, notando-se ainda que os mesmos estão em sequência e possuem aproximação fonética.

Desta forma, a relação entre o WER e o LD trará à tona os segmentos que apresentam maiores índices de erro e distância entre os caracteres. Serão analisados 406 segmentos de legenda que apresentarão os dois valores, WER e LD, para que se determine quais sequências irão para a análise qualitativa, de forma a identificar as estruturas linguísticas necessárias para o treinamento de *software*, para que ocorra o aprendizado e as legendas fiquem mais precisas, concluindo o objetivo desta pesquisa. O mesmo algoritmo criado para a análise de WER foi utilizado para o cálculo de LD.

3.4.4. Procedimento d – Levantamento dos índices de WER

O ponto central da análise quantitativa é estimar estatisticamente o quanto o *software* está errando através da métrica WER e a distância entre o que foi gerado automaticamente e o conteúdo ideal. Neste momento da análise quantitativa, busca-se levantar os segmentos com porcentagens de erros acima e abaixo de 50%, pois considera-se neste momento apenas os índices WER de cada segmento, dividindo-os entre positivos (até 49%) e negativos (de 50 a 100%). Desta forma, o quadro 6 demonstra como exemplo dois segmentos com pontuações extremas, sendo um com 100% de WER e o outro com 0%.

Quadro 6 - Exemplo de uma pontuação alta e uma pontuação baixa na análise WER

N. Segmento	Transcrição Automática	Transcrição ideal	WER	WER %
225	ocorrendo pela oferta da aquicultura	ocorrendo pela oferta da aquicultura.	0	0%
322	em 2008/2009 esse consumo aparente pulo para um milhão e	esse consumo aparente pulou para	1,00	100%

Fonte: Elaboração da autora.

O Quadro 6 mostra os dois extremos no cálculo de WER, o 0% e o 100%. Porém, na seção de resultados serão discutidos e demonstrados também os erros intermediários (0-25%, 25-50%, 50-75% e 75-100%), embasando a análise qualitativa no que diz respeito ao procedimento dos tipos de erros propostos.

Neste caso, demonstrado no segmento 225, há apenas uma diferença do ideal em relação ao automático, o ponto final da sentença, ou seja, a LD para este segmento foi igual a 1, embora o WER tenha apresentado 0%. Questões de pontuação como estas serão discutidas nas análises qualitativas.

3.5. Procedimentos de Análise qualitativa de transcrição automática

Conforme exposto anteriormente, esta pesquisa leva em conta dados coletados de forma quantitativa com a aplicação das fórmulas WER e modelos comparáveis através da equação de LD. Tais inferências indicam o quanto o *software* foi eficiente, isto é, a quantidade de acertos *versus* a quantidade de erros com base na transcrição ideal que o *software* deveria ter feito. Com isso, é possível identificar, através das métricas WER e LD, além do erro total (em todo o recorte analisado), o quanto cada segmento de legenda atingiu de erro.

Porém, tanto a porcentagem de erros estimadas no processo quantitativo, quanto a distância de caracteres entre palavras que ocorrem no segmento ideal, não são suficientes para determinar a qualidade do que está sendo gerado pelo *software*, nem mesmo a determinação dos modelos de treinamento propostos. Mesmo que os dados quantitativos se mostrem positivos, só é possível determinar, de fato, quais são as problemáticas do processamento na execução de uma análise qualitativa comprovando os elementos quantitativos (LIMA-LOPES; MORO, 2018, p. 62-63). Essa necessidade da análise qualitativa se faz presente, pois, mesmo analisando a distância entre os caracteres do automático e do ideal, não é possível inferir a que se decorreu a mudança daquela sentença ou palavra. Assim como não há uma forma de padronizar os tipos de erros ocorridos se não for por uma análise manual.

Trata-se, assim, de uma observação do co-texto cujo objetivo é fornecer insumos para o treinamento. Ao analisar os processos da linguagem da máquina, o conhecimento dos contextos de cultura e situação estão sendo levantados nesta pesquisa (HALLIDAY, 2005).

Portanto, com relação à 2) análise qualitativa, foram executados os seguintes passos, a) levantamento dos tipos de erro ocorridos levando em conta as porcentagens mais altas e baixas no cálculos de WER, b) análise de palavras-chave[35] e suas ocorrências dentro das sentenças, analisando o contexto com relação ao tipo de erro em que ela foi atribuída e c) a proposição dos modelos de treinamento do *software*.

---

[35] Palavras-chave: palavras que determinam o contexto da frase e a inter-relação entre o conteúdo exposto.

3.5.1. Procedimento a - Levantamento dos tipos de erro ocorridos

O primeiro procedimento qualitativo a ser executado é o levantamento dos tipos de erros ocorridos nos segmentos analisados. Estes tipos de erros darão o embasamento para verificar a origem das porcentagens dos cálculos de WER e as pontuações de LD. Para tanto, será possível efetuar a separação dos modelos de treinamento da *Skylar*.

Os tipos de erros apresentados no Quadro 7, abaixo, foram levantados manualmente durante o processo de PE do *corpus* e serão analisados de forma qualitativa na seção de resultados, levando em conta os cálculos quantitativos. Os erros encontrados foram classificados em 3 grupos de erro para análise, sendo Tipo 1 – Leve, Tipo 2 – Médio e Tipo 3 – Grave.

Quadro 7 - Tipos de erros do corpus analisado

Tipo de erro	Classificação	Descrição	Exemplo
Tipo 1	Leve	- Alterações optativas por segmentos fora dos padrões de legendagem. - Alterações de repetição de palavras pelo locutor, interjeições ou palavras ditas de forma incorreta.	- Seg. 171 Automático: e aqui se a ideia é olha **o que que** me motivou a Ideal: E aqui sim**.** A ideia é**: "**Olha o que me motivou a*
Tipo 2	Médio	- Palavras com proximidade fonética.	- Seg. 391 Automático: eu **cansei** dois cenários por conta dessa diferença Ideal: Eu **tracei** dois cenários por conta dessa diferença,
Tipo 3	Grave	- Ausência de palavras; - Palavras com sentido oposto; - Nomes próprios interpretados como outros substantivos; - Palavras aleatórias.	- Seg. 326 (ausência de palavras) Automático: foi consumido fora Ideal: foi consumido fora **de casa**? - Seg. 256 (palavras com sentido oposto) Automático: quanto algo em torno de quase **3 milhões** de reais aula Ideal: Grande quanto? Algo em torno de quase **3 bilhões** de reais ao ano, tá? - Seg. 282 (nomes próprios como substantivos) Automático: através do IBGE **bebê** já tem uma pesquisa que chama Ideal: através do IBGE. O **IBGE** tem uma pesquisa que chama** - Seg. 272 (palavras aleatórias) Automático: **colchas** em relação aos dados ela surgiu já há dois Ideal: **Como eu falei para vocês, minha desconfiança** em relação aos dados,

* e **: Alguns segmentos apresentam mais de um exemplo de erro. Isso será tratado na seção de resultados.

Fonte: Elaboração da autora.

O erro tipo 1 – Leve, tem como tipos de erros duas alterações consideradas optativas no processo de PE. No que diz respeito às alterações optativas por segmentos fora dos padrões de legendagem, leva-se em consideração a pontuação, disposição da sentença. Este primeiro tem conexão com o segundo tipo de erro apresentado, a repetição de palavras pelo locutor. A pontuação não é considerada no processo de transcrição do *software*, por isso foi inserida como um erro Leve, já que a maioria dos segmentos irá apresentar um erro deste tipo. Além disso, a pontuação e disposição do segmento fazem parte dos critérios de legendagem que devem ser seguidos na elaboração dos trechos.

Quanto à repetição, é comum que ocorra na fala em algumas palavras, como no exemplo: a repetição do "que", é uma maneira comum de se falar de forma coloquial, portanto também foi atribuída ao tipo de erro Leve. Este tipo de erro não compromete o entendimento do receptor da mensagem, caso ele consumisse apenas o conteúdo automático. Então, embora o *software* tenha transcrito da maneira correta todas as palavras ditas pelo locutor, não significa que seria o ideal a ser apresentado.

O erro tipo 2 – Médio, tem como tipo de erro a proximidade fonética das palavras. Este é um erro muito comum em todo o *corpus*, mas que pode comprometer moderadamente o entendimento do receptor que consumirá o conteúdo audiovisual com estas legendas, visto que a similaridade fonética gera palavras incorretas. No caso apresentado no quadro, o segmento, além de ser interpretado pela máquina com uma palavra de outro sentido (considerado também erro tipo 3 – Grave), apresenta a proximidade fonética no sufixo das palavras "cansei" e "tracei". Toda a frase foi comprometida por este equívoco, porém pelo contexto nos segmentos seguintes foi possível compreender o sentido do que estava sendo dito.

Por fim, o erro tipo 3 Grave, contempla três categorias de erros importantes para o entendimento do trecho. O primeiro tipo tem relação com a ausência de palavras, ou seja, o conteúdo não é transcrito na íntegra. Como mostra o exemplo, no trecho automático foram ocultadas as palavras "de casa"[36], crucial para o entendimento do conteúdo do professor. O segundo trecho apresenta palavras com sentido oposto, que podem ser provenientes de um erro de proximidade fonética (tipo 2), que geralmente aparecem nos mesmos trechos. No exemplo, o professor se referia a "3 bilhões de reais" e não a "3 milhões" como a transcrição automática sugeriu. O terceiro tipo se refere à interpretação de nomes próprios como substantivos comuns. Este erro ocorre, pois, o processamento do reconhecimento de voz busca o padrão nas palavras mais faladas, sendo um erro mais difícil de tratar, uma vez que nomes próprios podem se manifestar de várias maneiras sem relação com palavras existentes. No exemplo, o *software* interpreta corretamente o termo "IBGE", porém, quando há a repetição deste, ele entende como a junção dos dois Bs (dito de forma bem enfática pelo professor[37]), transformando o tempo

[36] Os dois termos também não aparecem nos segmentos da sequência da transcrição.
[37] Pode-se conferir o vídeo no anexo 1.

no substantivo “bebê”. Por fim, o quarto tipo apresenta palavras aleatórias sem nenhum sentido com a frase ou com a fonética do que foi dito no momento. Como podemos ver no exemplo, a frase “Como eu falei para vocês, minha desconfiança em relação aos dados,” foi interpretada por uma única palavra aleatória: “colchas”.

É com base nestes grupos de erros que o primeiro procedimento da análise qualitativa se iniciará, sendo elaboradas interpretações e levantamentos dos tipos diferentes de erros em cada grupo para a elaboração dos modelos de treinamento.

CAPÍTULO 4 - RESULTADOS E DISCUSSÃO

4. Do *corpus* de pesquisa

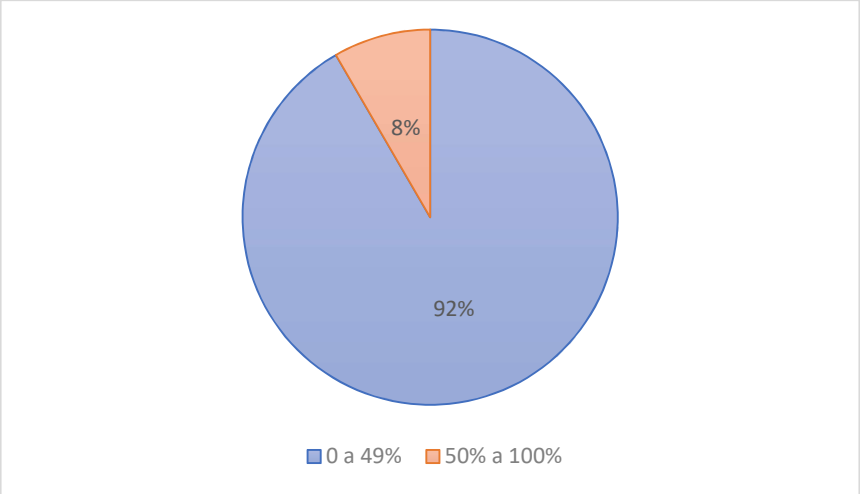
Como já colocado anteriormente, esta análise parte de um *corpus* no formato de legendas elaborado pelo *software Skylar*. As legendas foram geradas pelo *software* e retiradas da videoaula "Agronegócio do Pescado" oferecida pelo Pecege no curso de Agronegócios. A aula tem, ao todo, 2 horas e 28 minutos, o que gerou 1658 segmentos de legenda. Para a análise desta pesquisa, foram coletados 406 segmentos ao todo, totalizando 31 minutos de aula (Anexo 2). Todos os dados de "transcrição ideal" foram gerados manualmente através do procedimento metodológico de PE descrito na seção de Metodologia. Os cálculos de WER e LD foram gerados através de um algoritmo desenvolvido especialmente para esta pesquisa dentro da própria ferramenta analisada, a *Skylar*.

4.1. Análise quantitativa - Geral

Foram aplicados os cálculos de WER e LD em todos os 406 segmentos utilizados como análise através de um algoritmo desenvolvido especialmente para esta pesquisa, dentro da própria ferramenta.

Para iniciar as análises e resultados obtidos com a aplicação das métricas quantitativas, será estabelecida uma divisão entre dois grupos maiores considerados "positivos" e "negativos". Para os positivos, serão considerados os segmentos que obtiveram porcentagem WER igual ou menor a 49%. Para os negativos, segmentos que obtiveram porcentagem WER igual ou acima de 50% (Gráfico 1).

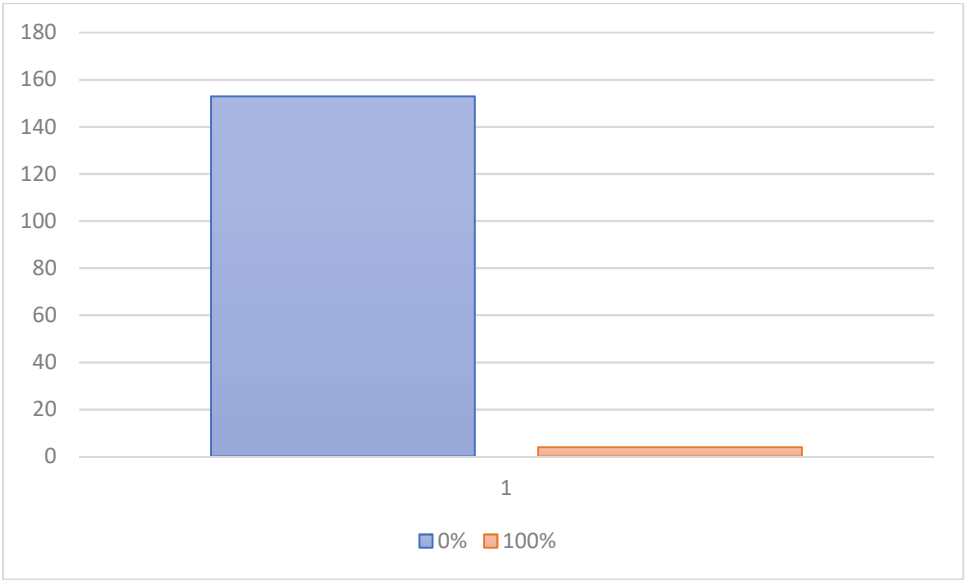
Gráfico 1 - Relação das extremidades do cálculo WER



Observa-se no Gráfico 1 que, dos 406 segmentos analisados, 372 obtiveram um índice de erro WER inferior a 49%, totalizando 92% do recorte. Ou seja, dos segmentos que foram positivos em sua execução automática, poucas alterações em quaisquer classificações apresentadas serão tratadas posteriormente na análise qualitativa.

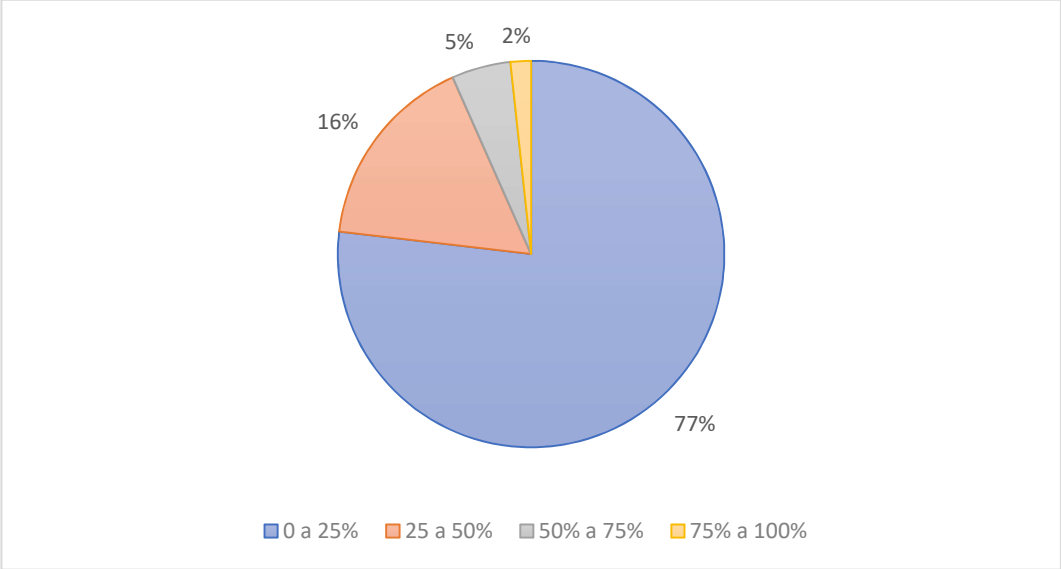
Já no que diz respeito ao segundo grupo que considera as porcentagens de 50 a 100%, apenas 34 segmentos representam um resultado significativo no âmbito negativo, sendo: 8% da coleta – segmentos que apresentam grandes problemas estruturais quando comparado o processo automático e ideal. O Gráfico 2 mostra a relação entre segmentos com 0% e 100% do recorte analisado.

Gráfico 2 - Resultado de segmentos que totalizaram 0% e 100% de WER



Dentro dos segmentos considerados positivos, foram contabilizados 153 segmentos representando 0%, ou seja, não houve alterações nas palavras do trecho. Porém, isso não significa, necessariamente, que nenhuma alteração ocorreu, como veremos discussão qualitativa através do índice de LD, que nos mostrará a pontuação de alterações. Mesmo assim, com alterações que entraram na classificação optativa, estes dados mostraram o melhor desempenho do recorte até então, em termos quantitativos. Nos segmentos negativos, apenas 4 segmentos obtiveram WER 100% na relação automático x ideal. O Gráfico 3, por sua vez, apresenta o demonstrativo separado em 4 grupos de porcentagens, variando de 0-25%, de 25 a 50%, de 51-75% e de 76-100%.

Gráfico 3 - Relação dos grupos de WER em todo o corpus



Nota-se pelo Gráfico 3 que a maior indicativo como porcentagens WER baixas (na série de 0 a 25%) concentra-se em 77% do demonstrativo, indicando que, no geral, o *software* obteve um bom funcionamento quando comparado o automático e o ideal. O segundo maior indicativo, com 16% de ocorrência, recai sobre os erros na faixa de 25 a 50%, o que indica uma quantidade significativa de erros ainda dentro das porcentagens consideradas positivas para o processamento do *software*. As faixas negativas, acima de 50%, demonstram porcentagens baixas como expostas anteriormente.

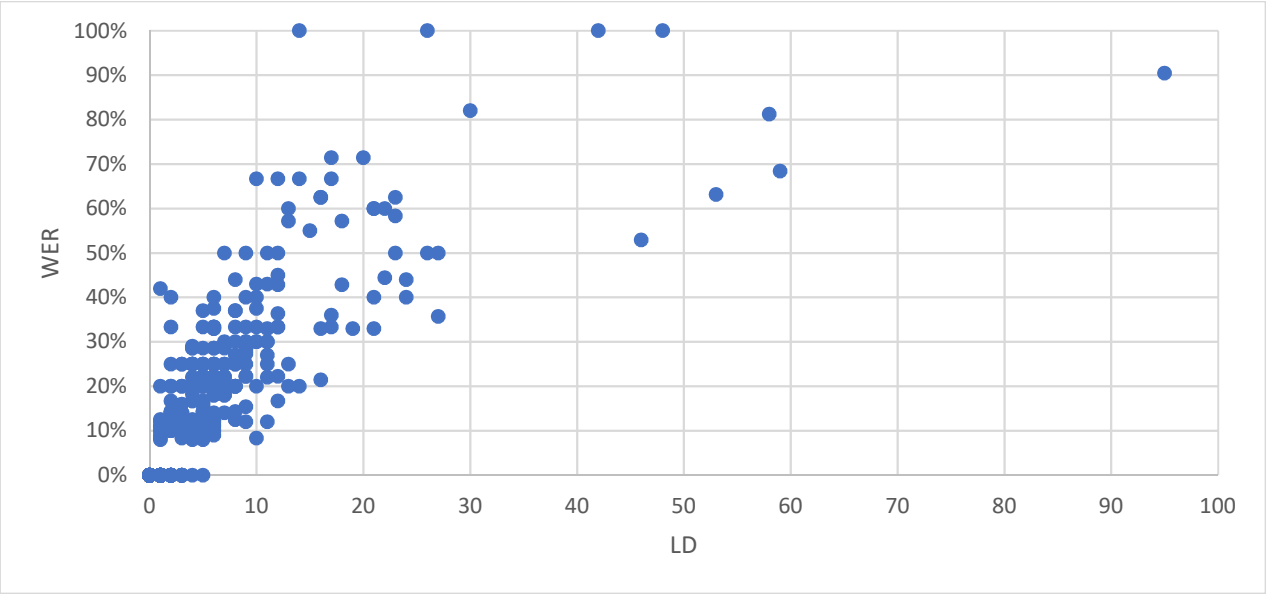
4.1.1. Análise quantitativa – com distribuição entre WER e LD

A partir das discussões das métricas WER e LD, separadas, este estudo também organizou os resultados combinados dessas estatísticas, gerando uma melhor visualização de como elas podem expressar aqueles segmentos de legenda que tenham sido diferentes do ideal.

Uma análise preliminar pode pressupor que estas métricas estejam diretamente relacionadas. Contudo, esta seção realiza a aferição desta correlação, bem como fornece o entendimento da influência dessas métricas no erro de cada segmento.

No Gráfico 4, portanto, observa-se a dispersão entre os resultados de WER e LD para cada um dos segmentos do recorte deste trabalho. Cada ponto do gráfico representa um segmento da amostra analisada.

Gráfico 4 - Gráfico de dispersão entre WER e LD



Como pode-se observar, a maior parte dos segmentos esteve perto do ponto origem do gráfico, no quadrante com LD menor que 20 caracteres e WER menor do que 30%. Esta análise infere que a maior parte dos segmentos apresenta um erro baixo, conforme a análise da seção anterior permitiu afirmar. Ainda, pode-se notar alguns pontos distantes de tal quadrante, evidenciando segmentos que gerariam quase nenhum entendimento para o leitor por apresentarem todas as palavras diferindo do ideal –

impacto que será analisado e apresentado na seção de análise qualitativa. Por fim, nota-se que não tivemos segmentos com WER baixo (menos do que 40%) e LD alto (maior do que 40). Tal fato ocorre por um LD alto, em geral, indicar que muitas palavras foram substituídas, inseridas ou removidas, o que também acarretaria um WER alto. Já o oposto pode não ser verdadeiro, ou seja, pode-se ter um segmento com um LD baixo (poucos caracteres trocados), mas que impactem em muitas (ou até mesmo todas) palavras diferentes da transcrição ideal.

Esta análise também foi expandida para histogramas individuais das métricas WER e LD, conforme mostram os Gráficos 5 e 6. Essas ferramentas permitem uma melhor visualização da concentração das ocorrências de ambas medidas, a fim de embasar a afirmação de poucos índices altos de erros individuais.

Gráfico 5 - Histograma de LD

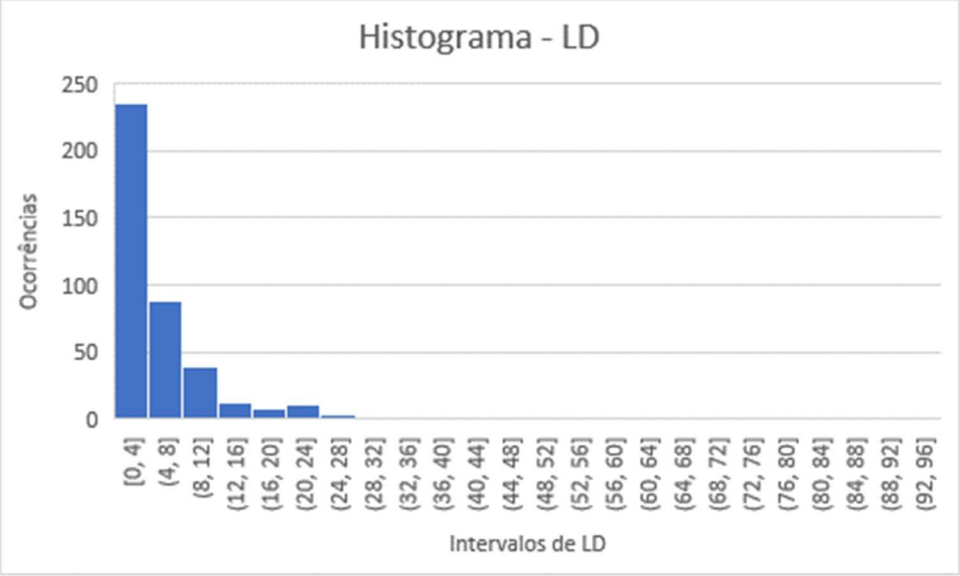
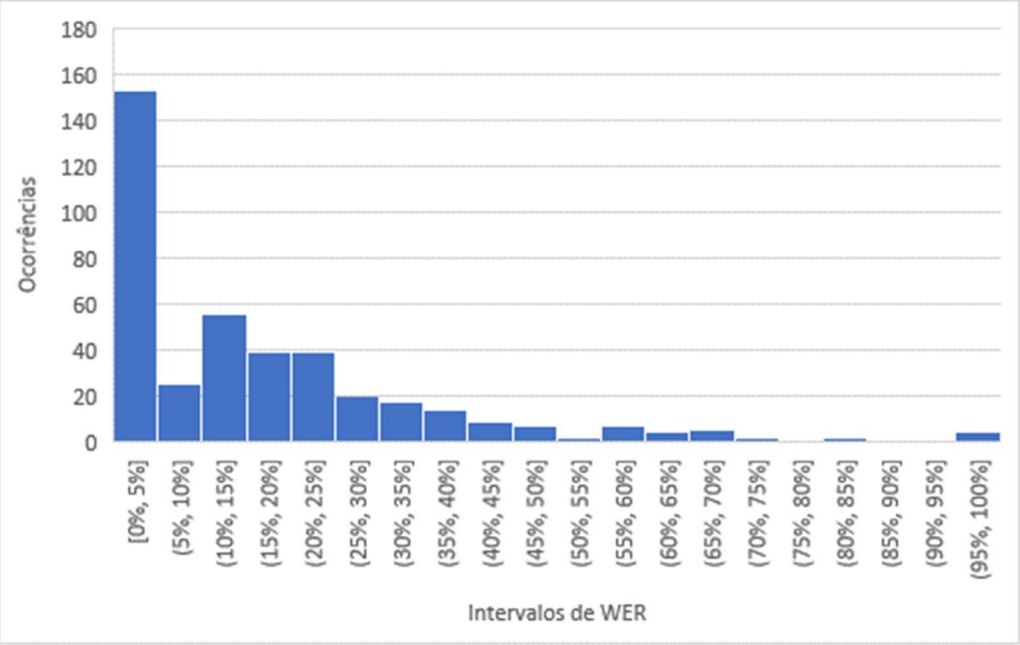


Gráfico 6 - Histograma de WER



A partir dos gráficos acima, pode ser observado que, de fato, a maior parte dos erros de LD e WER foi baixa, com destaque para a tendência de decréscimo das ocorrências *versus* o aumento dos intervalos de erro. Nota-se, ainda, que para a métrica WER, a maior parte dos segmentos teve índice de erro menor do que 5%, porém o número de ocorrências de WER entre 5 e 10% foi menor que o total de ocorrências para seus 3 intervalos subsequentes – entre 10 e 15%, entre 15 e 20% e entre 20 e 15% - essas ocorrências mostram que, apesar de na maior parte dos casos os erros estarem próximos de 0, ainda existe um número considerável de segmentos que apresentam erros de nível médio de ocorrência, isto é, entre 10 e 25% de WER.

Este trabalho ainda propôs uma análise da temporariedade das ocorrências de erros no decorrer da aula analisada, ou seja, em quais momentos da aula existe maior frequência de erros. Para tanto, os gráficos 7 e 8 evidenciam as ocorrências desde o segmento 1 até o segmento 406, para WER e LD, respectivamente.

Gráfico 7 - WER para cada segmento

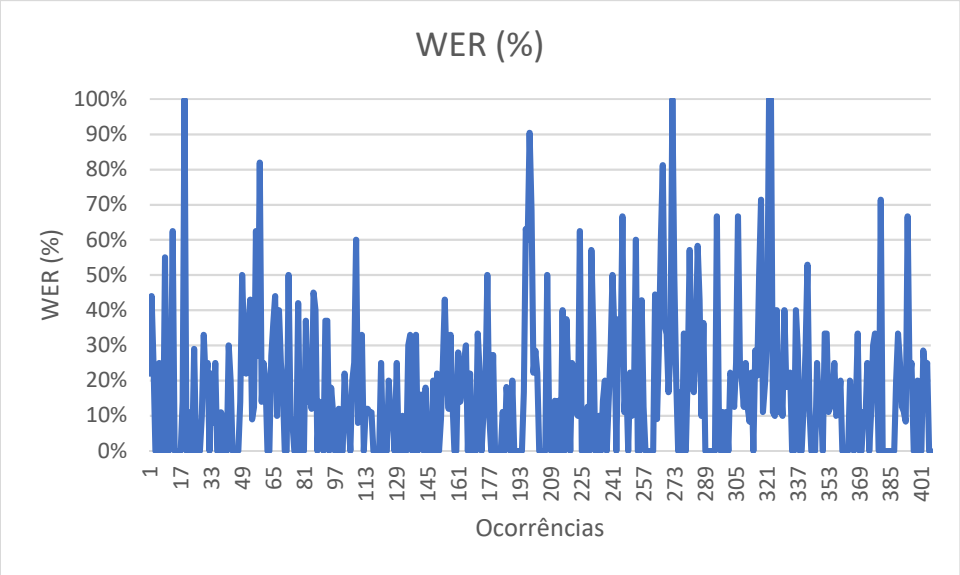
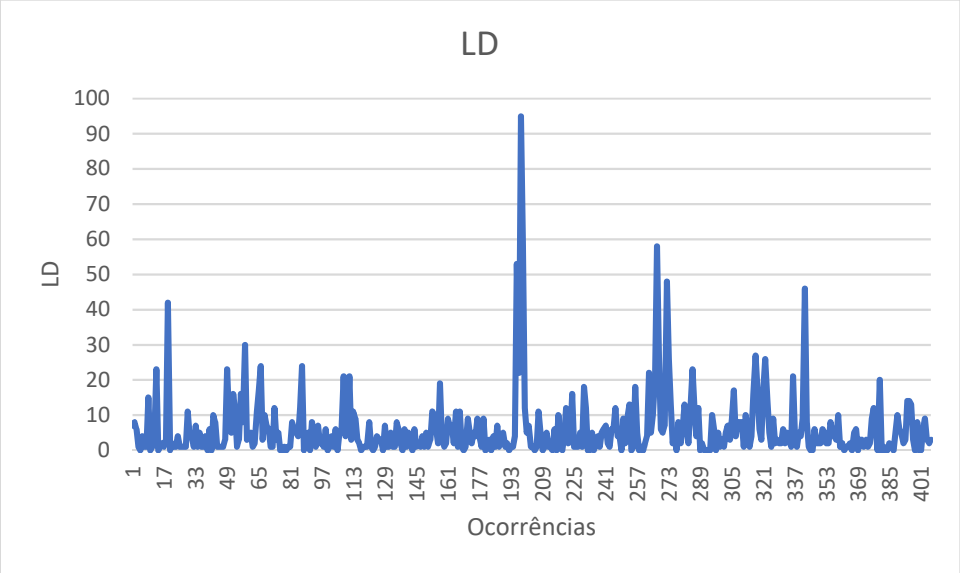


Gráfico 8 - LD para cada segmento



Conforme nota-se, não existe uma relação específica direta entre o momento da aula e a quantidade de erros. De outro modo, destaca-se uma possível dificuldade a qual o *software* apresentou ao identificar as palavras por motivos variados listados na seção qualitativa (4.2) A premissa também se faz verdadeira para a métrica LD, observando-se que não existe qualquer condição de momento da aula que aponte para mais erros em tal momento. Para a LD, nota-se alguns segmentos específicos que apresentaram erro grave os quais devem ser analisados individualmente para propor-se uma correção.

Com isso, conclui-se a análise quantitativa do recorte aplicado neste trabalho. Como conclusões gerais pode-se identificar que, apesar de o *software* acertar a maior parte dos segmentos, ainda existe um espaço para a identificação dos mesmos de uma forma mais apurada, buscando a redução dos erros de transcrição da ferramenta, que é objetivo geral deste trabalho.

Destaca-se ainda que as métricas WER e LD foram ferramentas importantes nesta análise, por permitirem quantificar o quanto o *software* está acertando e errando e, ainda, através de análises aplicadas a essas medidas, permitir isolar os maiores erros a fim de analisa-los para cumprir o objetivo deste estudo.

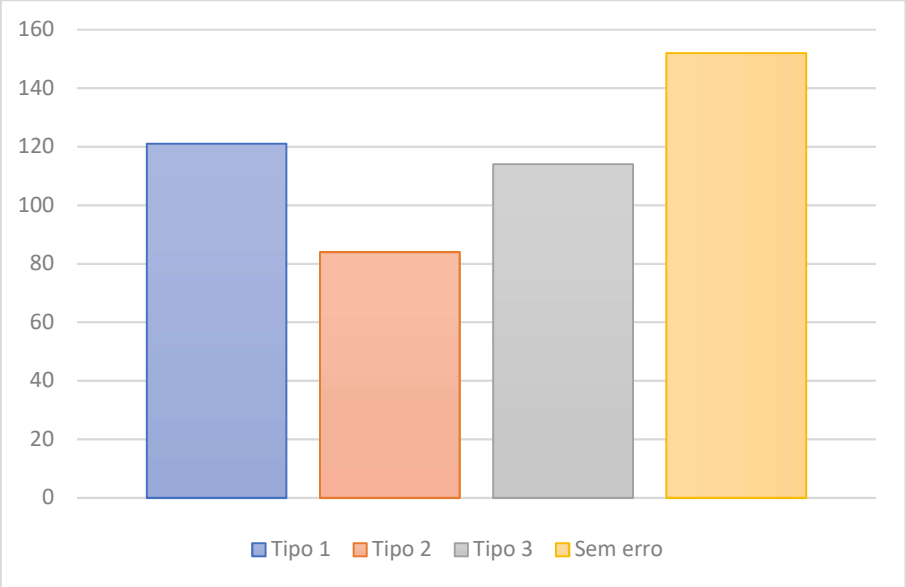
Com base nisso, a próxima seção realiza a discussão qualitativa dos segmentos aqui analisados.

4.2. Análise qualitativa - geral

Para a análise qualitativa, foram utilizados como base os dados quantitativos da seção anterior, elencando alguns pontos essenciais para a elaboração dos modelos de treinamento linguísticos do *software*, objetivo primeiro desta pesquisa. Essa subseção se dedica à exploração dos dados gerais e procura explicar como os erros foram atribuídos e a classificação dos segmentos. Assim, será apresentado um panorama geral que servirá, posteriormente, como base de comparação entre os dados quantitativos levantados.

Foram analisados os 406 segmentos de forma manual, sendo feita a classificação de cada tipo de erro apresentado nos segmentos (visualização no anexo 3). O Gráfico 9 apresenta a relação entre os segmentos com a divisão entre erro tipo 1 - leve, tipo 2 – médio, tipo 3 – grave e os segmentos sem erros. Durante a aferição dos dados, alguns segmentos apresentaram mais de um tipo de erro, ocasionando em um número final de 319 segmentos atribuídos com ao menos um dos tipo de erro considerados e 152 segmentos aos quais não foram atribuídos erros. No total, foram geradas 471 análises.

Gráfico 9 - Distribuição dos erros entre os tipos

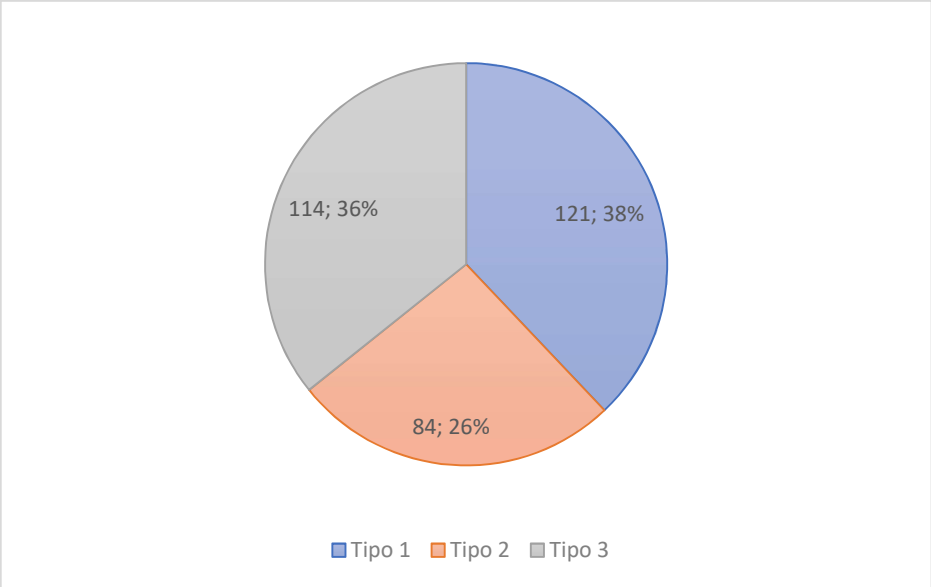


O gráfico acima mostra a distribuição entre os segmentos com erros e sem erros apurados a partir do Quadro 7 – relação dos tipos de erros apresentada na seção da Metodologia. Para iniciar esta seção qualitativa, usa-se ainda dados quantitativos embasando os segmentos que serão analisados de forma mais específica. Como pode-se

observar, a coluna “sem erro” apresenta um maior índice em relação àquelas que possuem algum tipo de erro atribuído, porém não se vê uma disparidade em relação aos segmentos que foram atribuídos um tipo de erro ou mais, como foi observado nos gráficos quantitativos. Isso nos indica que, observando apenas os dados quantitativos, a quantidade de segmentos apresentados com o índice WER baixo e LD também baixo, isto é, todos esses segmentos estejam corretos e não possuam erros relevantes, o que já inicia essa seção de forma notável.

Ainda a respeito do Gráfico 9, são apresentados dados em que se tem algum tipo de erro atribuído, por ordem de classificação: o tipo 1, tipo 3 e tipo 2. O Gráfico 10 demonstra a quantidade de erros que foram distribuídos em uma amostra de 319 erros atribuídos.

Gráfico 10 - Distribuição por porcentagem de segmentos com erros atribuídos



O Gráfico 10 apresenta a quantidade de erros atribuídos e a distribuição entre eles, conforme dito anteriormente. Como pode-se observar, o erro do tipo 3 – grave, aparece com apenas 2 pontos abaixo do erro tipo 1 – leve, fator que será desenvolvido nas próximas seções onde se tratará das questões específicas de cada atribuição dos erros. O erro do tipo 3 é considerado grave pois pode ocorrer com os três itens mais problemáticos para a transcrição: falta de palavras transcritas, nomes próprios reconhecidos como substantivos comuns e palavras de sentido oposto. Durante a avaliação e classificação dos

erros, notou-se que alguns deles ocorriam em ao invés de outros, por exemplo, erros de proximidade fonética causavam erros graves de alteração de sentido ou interpretação errônea das palavras processadas. A proximidade dos erros de tipo 1 e 3 foram ocasionais, não havendo relação entre eles. Já os tipos 2 e 3 possuem uma relação quase que direta, fator a ser discutido posteriormente.

Levando-se em conta a aferição dos dados acima, as próximas seções tratarão especificamente dos exemplos extraídos de cada tipo de erro e será apresentada a análise dos motivos pelos quais alguns tipos de erros foram atribuídos.

4.2.1. Análise qualitativa – por tipos com exemplos

Durante a análise dos dados comparando a transcrição automática e a ideal, também levando em conta os cálculos WER e LD apresentados (como uma forma comparativa entre os dados quantitativos e qualitativos), foram atribuídos os tipos de erros aos segmentos. A análise a ser demonstrada nessa subseção tem como objetivo separar segmentos significativos para a construção dos modelos, sendo ordenados pelos grupos de segmentos positivos (até 49% de WER) e negativos (de 50 a 100% de WER), comentando os tipos de erros atribuídos a eles. Além desta análise, serão enfatizados alguns segmentos com exemplos mais críticos em relação aos erros, que não entraram nas relações dos segmentos por % de WER. Por fim, serão comentados exemplos específicos em relação ao comportamento do discurso do professor, a fim de justificar o porquê de alguns erros e mudanças necessárias ocorrentes.

Os quadros de análises a seguir foram compilados com os erros do tipo 1, 2 e 3 (ver seção Metodologia), seguindo a ordem do WER da menor para a maior porcentagem. Nestes quadros estarão apenas alguns recortes dos segmentos para discussão e os quadros com todos os segmentos positivos e negativos estão disponíveis para consulta no Anexo 3 e 4, respectivamente.

4.2.2. Discussão dos tipos de erros encontrados com mais frequência

O Quadro 8 apresenta os segmentos com erros do tipo 1, o primeiro a ser discutido. Conforme exposto anteriormente, erros do tipo 1 são considerados leves, pois possuem alterações optativas para o encaixe no padrão de legendagem e remoção de repetições ou interjeições ditas em excesso pelo professor, não interferindo necessariamente no entendimento do segmento.

Quadro 8 - Exemplos de segmentos com as classificações de erro tipo 1

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
186	água **ela** é regulada por esse Ministério Ministério da Pesca e	A água é regulada por esse Ministério, Ministério da Pesca e	7	18%	1
61	mente o seguinte o Brasil hoje **ele** é um grande	mente o seguinte: o Brasil hoje é um grande	5	20%	1
3	prazer estar aqui novamente com vocês hoje substituindo	É um prazer estar aqui novamente com vocês hoje, substituindo	6	25%	1
52	fontes de dados principais fontes de dados que nós	principais fontes de dados que nós	16	33%	1
66	pequena a principal a principal a principal fonte de	pequena. A principal fonte,	24	44%	1
49	**que** nós já tivemos **aí em situações** de **em** situações	nós já tivemos de situações	23	50%	1
295	seria muito mais elevado do que **é o que é**	seria muito mais elevado do que	10	67%	1
306	o consumo **da pop** deu 774 e a minha justificativa	deu 774 e a minha justificativa	17	67%	1

Como pode-se observar, a maioria dos erros encontrados foram classificados como tipo 1, considerando-se uma inadequação do tamanho dos segmentos, de acordo com as regras de legendagem (NAVES *et al*., 2016). Tal inadequação se manifesta, principalmente, pela inclusão de itens lexicais que poderiam ser omitidos. Nos seguimentos 186, 61, 49 e 295 o processo de transcrição manteve pronomes pessoais (ela, ele) e pronomes relativos (que). Já em seguimentos como 306 e 49, tais erros se manifestam na inserção de substantivos e advérbios, respectivamente.

No segmento 186, observa-se que o pronome pessoal *ela* é repetido enfatizando-se o objeto para o qual está direcionando seu discurso, ao passo que, no segmento 61, o

professor utiliza o pronome “ele”de forma coesiva, uma catáfora para se referenciar a algo que já havia definido anteriormente “o Brasil”. Tais estratégias ocorrem em diversos momentos durante a aula e, apesar de sua importância para o para a interação com o aluno, elas não deveriam estar presentes nas legendas finais.

No segmento 3 (quadro 8), observa-se que, com índice WER 25%, há apenas uma inserção “É um” que havia sido transcrita no final do segmento anterior e foi alterada para o segmento atual. Tal alteração ocorreu por uma opção de correção, sem influenciar o conteúdo, mas visando uma melhoria da compreensibilidade.

Há 3 casos de repetição no quadro 8 (segmentos 52, 66 e 295). No segmento 52, podemos notar uma repetição efetuada pelo professor no momento da fala, no caso, a expressão “fonte de dados”, que foi corrigida por “principais fontes de dados”. Tal inadequação ocorreu graças a um momento de hesitação aparente no início do segmento anterior, no qual o professor inicia falando sobre as “principais fontes de dados”, enfatizando o que estava mostrando em seus slides. O mesmo ocorre no segmento 66 com a repetição de “a principal”, na qual há a repetição intra-segmental. A hesitação também propiciou o erro no segmento 295, no qual há a repetição incorreta. Ela decorre de um marcador, de uma forma de expressão oral com o som de “é” no meio do segmento, sendo este o momento em que o professor está trocando os slides para demonstração dos dados da aula. No segmento 49, podemos notar também uma repetição da expressão “em situações”, novamente como forma de hesitação, porém há uma remoção executada no advérbio “aí”, caso também relevante para análise, uma vez que tornou-se recorrente ao longo da transcrição (ver 4.2.3).

Por fim, a 306 representa um caso pertinente para a classificação dos erros. Neste segmento, a palavra “pop” interpretada pelo processo automático é, na verdade “POF”, o que ocasionaria em um erro do tipo 3. Porém, foi erroneamente interpretado pelo *software*, uma vez que o contexto dessa fala ocorre no segmento 305 (Anexo 2 e 3). Portanto, foi removido deste segmento pois já havia sido dito em outro, justificando a classificação do tipo 1.

Foi observado que os erros não ocorrem de forma isolada nos segmentos, uma vez que a construção de cada trecho é elaborada a partir daquele momento de fala do

professor. O quadro 9 apresenta alguns segmentos em que houve erros do tipo 1 e 3 – graves.

Quadro 9 - Exemplos de segmentos com as classificações de erro tipo 1 e 3

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
64	doce ela ela é muito grande no Hemisfério Sul AC da	doce é muito grande no Hemisfério Sul aqui da	11	27%	1,3
42	durante todo o c* todo esse momento que vou ficar	durante todo esse momento que vou ficar	10	30%	1,3
207	**e**xpectativa de todos **aí** principalmente daqueles que	**Então** a expectativa de todos, principalmente daqueles que	11	50%	1,3
273	anos atrás e ela surgiu em função dessa tabela aqui	surgiu já há dois anos atrás e em função dessa tabela aqui.	26	50%	1,3

No caso do segmento 42, a transcrição foi executada de forma correta, porém, por um momento de hesitação ao dizer a palavra "curso", o que a deixou incompleta,. Nesta, o *software* entendeu como "c*", no caso, "cu" e foi bloqueada com filtro de palavrões já desenvolvido previamente. Embora o erro tenha sido leve, levou a classificação de erro 3 por ter sido inserida uma palavra aleatória e de baixo calão. Já no segmento 64, o professor fala "aqui da" e o *software* reconhece como a sigla "AC", pressupondo, pela fonética, que era essa sigla representada no discurso. Isso pode ocorrer pelo fato de o *software* não compreender a palavra em si, mas os caracteres que são construídos para formas as palavras (ver Capítulo 2).

No caso de 207, a palavra "expectativa" foi processada de forma incompleta pois a fala ficou ligeiramente mais rápida e com um tom elevado de um segmento para o outro, causando uma desordem no processamento, o que ocasionou em um erro do tipo 3 por falta de um caractere em uma palavra. Também neste segmento, foi removido o advérbio "aí". Durante toda a transcrição, as duas palavras mais repetidas pelo professor foram os advérbios de lugar "aí" e "aqui". No caso especificamente do "aí", o professor utiliza de duas maneiras, tanto como uma indicação de lugar, quanto como interjeição no meio das frases.

Por fim, o caso 273 demonstra um momento de hesitação na fala, ocasionando uma desordem no processamento. No segmento anterior a esse já haviam sido repetidas algumas palavras do trecho.

A maioria das ocorrências observadas derivaram de erros por proximidade fonética (tipo 2 - moderado). Tais erros, porém, raramente ocorreram de forma isolada, como pode-se notar pelo quadro 10 com 3 segmentos em destaque. Este resultado é determinante no que diz respeito à construção dos modelos de treinamento, pois infere-se que os erros fonéticos são comuns aos outros dois tipos de erros apresentados, se tornando um dos motivos principais pelos índices de erros do processamento. As combinações com os erros do tipo 2 serão apresentadas nos quadros 11 e 13.

Quadro 10 - Exemplos de segmentos com as classificações de erro tipo 2

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
6	surgirão então que envia as suas dúvidas e	sugiro então que enviem as suas dúvidas e	4	25%	2
35	agronegócio Pão Com certeza em Manaus pescado é	agronegócio? Bom, com certeza em Manaus pescado é	5	25%	2
73	nas **empresas represa** hidrelétric**a** e temos **fios** muito	nas represas hidrelétricas e temos rios muito	12	50%	2

O segmento 6 possui duas ocorrências de proximidade fonética: e “sugiro” para “surgirão” e “enviem” para “envia”. No caso do “surgirão”, por uma fala acelerada, a palavra foi processada juntamente com a palavra seguinte “então”, provocando uma confusão entre elas no momento da transcrição. No caso do “envia”, pelo mesmo motivo, houve alteração apenas em sua conjugação, não causando grande impacto no que foi transcrito. Os dois casos não foram caracterizados como tipo 3 por não comprometerem o entendimento do receptor. Já o segmento 35, podemos notar que a palavra “Bom” foi substituída no automático por “Pão”. Este erro ocorreu pela aproximação dos fonemas “P” e “B”, que possuem pontos de articulação próximos, em associação com vogais nasais.

O quadro 11 demonstra os erros do tipo 2 e 3, de forma integrada, ou seja, o impacto dos erros do tipo 2 que ocasionam no tipo 3. Nota-se um recorte mais significativo com estes erros combinados, pois o segmento se tornaria incompreensível, mesmo apresentando baixa porcentagem de erro no WER.

Quadro 11 - Exemplos de segmentos com as classificações de erro tipo 2 e 3

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
313	200 e o consumo da **pop** me apontou a estabilidade	e o consumo da **POF** me apontou a estabilidade,	6	22%	2,3
403	2008/2009 eu vou ter aí uma diferença **coceira** nas	2008/2009 eu vou ter aí uma diferença, **considerando** as	9	22%	2,3
404	**pernas** de 265000 toneladas	**perdas** de 265.000 toneladas.	3	25%	2,3
161	**final Fallout** sobre uma proposta financeira quem	**final falar** sobre uma proposta financeira **que**, quem	9	28%	2,3
376	**Cidade** tava meio desgostoso com essa informação um	**esse dado** eu tava meio desgostoso com essa informação, um	9	30%	2,3
53	utilizamos para **fazer alisar** os estudos alguns	utilizamos para **realizar** nossos estudos. Alguns	11	43%	2,3
154	grande costuma Costa do Atlântico muito muito	grande costa, a Costa do Atlântico muito grande,	10	43%	2,3
285	pelo consumo da **Pop 100** 2003 2002/2003 **16** entender a	**E** pelo consumo da **POF** em 2003, 2002/2003, só para vocês **entenderem**	23	58%	2,3
58	então **levante** de carne já co**m** Tiago e a ideia é	Vocês **já devem ter tido aula** de carne com o Thiago. E a ideia é	30	82%	2,3
19	então eu sugiro que você já entr**a** nesse	Então, eu sugiro que você já entre nesse **site e faça os downloads dos materiais,**	42	100%	2,3

Pode-se notar. neste recorte de 10 segmentos, metade com WER abaixo de 49% e metade com WER acima de 50%. Todos esses trechos apresentaram problemas graves em relação à proximidade fonética das palavras ditas pelo professor. Isso mostra que, independente da estatística levantada, os segmentos possuem problemas graves de interpretação dos trechos envolvendo, por exemplo, inserção de substantivo comum no lugar de nome próprio (seg. 313), inserção de nome próprio no lugar de verbo (seg. 161), inserção de uma palavra só no lugar de várias outras (seg. 58).

Nos segmentos 313 e 285 ocorreu um erro que foi comum em todo o trecho em que o professor se referia à "POF[38]". Não é comum que *softwares* de reconhecimento de voz façam a distinção de nomes próprios e siglas. Portanto, ele sempre irá buscar palavras conhecidas por ele e aproximar a transcrição ao que foi dito. A sigla "POF" foi confundida com "pop", no exemplo, e o substantivo "porta", dependendo das palavras que vinham antes ou depois do termo. Porém, um fato a ser notado foi que, no segmento 293 (Anexo 2), o *software* reconheceu o termo já na transcrição automática, dado que este trecho estava acompanhado de algumas palavras-chave como "dados da POF de 2009", fazendo com que a integração do *software* com o mecanismo de pesquisa do Google foi eficiente para este caso. Pode-se observar que nos outros casos em que "POF" foi reconhecido como "pop" ou "porta", os trechos envolviam palavras como "consumo da" e "como a POF mostra", respectivamente, não sendo relacionado com o termo em questão[39].

Ainda envolvendo os nomes próprios, o caso inverso aconteceu no segmento 161. Ao repetir duas palavras com o mesmo som da letra inicial "f", o *software* processou o nome próprio "Fallout[40]" no lugar de "falar". Não há indícios do motivo pelo qual isso ocorreu, exceto pelo fato de similaridade fonética, por isso a caracterização do erro 3 também para este quesito de palavra aleatória.

Seguindo a análise, os segmentos 403 e 404 mostraram casos de substituição de uma palavra por outra sem relação de sentido, como foi o caso de "coceira" para "considerando" e de "pernas" para "perdas", respectivamente. Esses segmentos demonstraram uma inversão simples de palavras por conta de similaridade fonética. Os segmentos 376, 53, 154 e 58 apresentam a proximidade fonética envolvendo duas palavras. No 376, o correto seria "esse dado" e foi transcrito como "Cidade", que foi ocasionado, muito provavelmente, por uma pausa na fala vinda do segmento anterior, além da proximidade fonética entre as palavras. O mesmo ocorreu de forma inversa no segmento 53: de "fazer alisar" para "realizar". Já no caso do 154, houve uma hesitação ao se referir à "Costa do Atlântico", e o *software* processou a palavra "costuma", próximo

---

[38] POF: Pesquisa de Orçamentos Familiares. Fonte: https://ww2.ibge.gov.br/home/estatistica/populacao/condicaodevida/pof/

[39] Esta interação entre o *software* e o mecanismo de pesquisa surge para dar mais precisão às palavras da transcrição e será discutido em trabalhos futuros envolvendo sistemas de cognição dentro do processamento de linguagem natural, ou compreensão de linguagem natural.

[40] Um jogo de RPG. Fonte: https://www.ludopedia.com.br/jogo/fallout

à “Costa”, pois a palavra “uma” foi proferida pelo professor no meio da sua fala enquanto a definia. Por fim, no 58, além de constar a omissão das palavras ditas, o *software* resumiu toda a similaridade fonética a uma única palavra incorreta: “levante”, além da falta do artigo que se referia à “Thiago”.

Por fim, os segmentos 285 e 19 apresentam dois casos diferentes dos anteriores. No 285, houve a junção dos sons de “s” e “f” sequenciais em “POF”, “2002”, “2003” e “vocês”, ocasionando na inserção dos numerais “100” e “16” na transcrição automática. Como já dito anteriormente, o *software* busca por alguma correspondência na base de pesquisa da *internet*, ocasionando similaridades entre os conteúdos como, no caso, uma sequência numeral. Já no 19, houve também uma sequência de sons com “s” “nesse site e faça os downloads dos materiais”, ocasionando na omissão por falta de compreensão do *software*.

Geralmente o erro do tipo 3 ocorre em função do erro tipo 2. Com isso, o quadro 12 abaixo mostra os segmentos que foram classificados apenas como tipo 3, de forma comparativa ao quadro 11. Vale ressaltar que a classificação dos erros isoladamente como tipo 3 podem ter sido ocasionados por problemas fonéticos, mas a gravidade do erro foi intensificada deixando-os isolados como somente tipo 3 pois, este equívoco não influencia somente o trecho, mas todo o tema da aula.

Quadro 12 - Exemplos de segmentos com as classificações de erro tipo 3

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
24	agronegócio do Pescado com professor Daniel **sonora**	Agronegócio do Pescado com o professor Daniel **Sonoda**.	4	29%	3
107	**ues-redes** dentro desses reservatórios	**tanques-redes** dentro desses reservatórios.	5	25%	3
317	um probleminha certo o **problema que a gente se o**	um probleminha, certo? **Então, o problema que eu montei foi o seguinte.**	27	50%	3
246	comercial de pescados **a gente vai mostrar para**	comercial de pescados. **Eu mostrei para**	12	67%	3
267	**8 -** tudo que eu exporto então **eu tenho um gráfico**	**O consumo aparente é a produção mais tudo que eu importo menos** tudo que eu exporto.	58	81%	3
323	**ladas**	**1.400.000 toneladas**	14	100%	3

O quadro 12 expõe 6 erros, especificamente do tipo 3, não tendo a influência de nenhum outro tipo de classificação. O único segmento que foge a essa regra é o 24. Neste segmento o sobrenome do professor foi confundido com o substantivo comum “sonora”, causando um problema de identificação do professor.

Em todos os outros segmentos apresentados há ocorrências de omissão de palavras ou letras que impactaram diretamente a compreensão da legenda, uma vez que estavam incompletas. Esses erros não têm relação com nenhum outro descrito até então, e parecem ter sido aleatórios por falha de processamento do *software*.

O quadro 13 expõe os segmentos que possuem classificação nos três níveis estabelecidos. Ocorreram problemas estruturais de legendagem, similaridades fonéticas que ocasionaram em uma palavra com o sentido dispare.

Quadro 13 - Exemplos de segmentos com as classificações de erro tipo 1,2,3

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
250	função da **apicultura** mas sem dúvida nenhuma aqui ó	função da **aquicultura** mas, sem dúvida nenhuma aqui, olha,	9	22%	1,2,3
176	basicamente é **apicultura** certo então **aquicultura aí**	Basicamente é a **aquicultura**, certo? Então a **aquicultura**,	9	50%	1,2,3
241	distorcid**a a** os resultados também **eles** acabam sendo	distorcidas, os resultados também acabam sendo	7	50%	1,2,3
394	um erro que **é o que esse povo tem que comer o peixe**	um erro que é, esse povo que come peixe	14	67%	1,2,3
198	ligação do uso da água para **agricultura** no Estado de São Paulo	E estão sendo criados, estão sendo feitas as regulamentações estaduais para o uso da água, ou a regulamentação para a **aquicultura**	95	90%	1,2,3
272	**colchas** em relação aos dados ela surgiu já há dois	**Como eu falei para vocês, minha desconfiança** em relação aos dados,	48	100%	1,2,3
322	**em 2008/2009** esse consumo aparente **pulo para um milhão e**	esse consumo aparente pulou para	26	100%	1,2,3

Neste quadro é apresentado um erro muito comum durante toda a análise do *corpus*: o conflito fonético entre a palavra "aquicultura" – correta na maioria dos casos – com "agricultura" e "apicultura". Podemos ver nos segmentos 250, 176 e 198 as ocorrências em que houve esses conflitos. No caso do 250, a palavra "apicultura", que é relacionada à criação de abelhas, foi inserida no lugar de "aquicultura", relativa à criação de peixes, tema da aula. No segmento 176, a primeira ocorrência foi novamente confundida com "apicultura". Depois, foi inserido de forma correta "aquicultura", caso relevante na criação dos modelos. Já no 198, a palavra foi confundida com "agricultura", que é a atividade de cultura do solo, não envolvendo de nenhuma forma as a criação de animais marinhos. Esses três segmentos também apresentaram outros problemas como a omissão de palavras (198) e uso de expressões como "ó" (250).

Ainda com relação à omissão de palavras, além do 198, o segmento 322 também se apresenta de forma incorreta, podendo ser por uma velocidade inesperada na voz do locutor. O segmento 272, além de apresentar omissão, modificou o sentido da frase, ou

seja, acarretando na interpretação de uma sequência de 7 palavras para a palavra "colchas", sem nenhum indício de similaridade fonética ou outras questões já discutidas. Esse exemplo foi caracterizado como inserção de palavra aleatória.

O quadro 14 demonstra um fator alternativo ao que foi discutido até então. Alguns segmentos apresentaram índice WER e pontuação em LD, porém não foram classificados com tipos de erros específicos.

Quadro 14 - Exemplos de segmentos sem classificações de erro com % de WER

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
187	aquicultura e também por mais seis órgãos Ministério do	Aquicultura e também por mais seis órgãos: Ministério do	1	11%	
189	superintendência de patrimônio da União Ibama e os	a superintendência de patrimônio da União, o Ibama e os	5	20%	

Em ambos os casos, todas as palavras exatas foram transcritas pelo *software*. Porém, as alterações de maiúsculas, por conta dos nomes próprios e pontuações por conta da sequência do que estava sendo dito, pontuaram para o índice WER, ainda que a frase não estivesse incorreta em termos de transcrição[41]. O índice WER levou em conta a inserção de caracteres de pontuação que foram descartados nessa pesquisa.

Esta seção qualitativa teve como objetivo elencar os principais tipos de erros que ocorreram no *corpus* transcrito, coletando os fatores que serão levados em consideração na elaboração de modelos de treinamento linguístico para a *Skylar*.

A próxima seção se dedica à divisão dos discursos durante a aula e as ocorrências similares de palavras e registros que ocorreram durante o discurso do professor. Para isso, será utilizado o *software* de LC, Sketch Engine.

[41] Ocorreram outros casos semelhantes, mas não há um recorte significativo para discussão neste trabalho.

4.2.3. Análise de palavras-chave e suas ocorrências

Essa seção se dedica à exploração das análises do *corpus* apresentado, com o objetivo de expor as palavras-chave e suas ocorrências dentro das sentenças. Para esse caso, será utilizada a ferramenta de análise de *corpus*, Sketch Engine. Além disso, faremos algumas comparações em relação ao *corpus* automático e ao ideal, buscando entender as diferenças entre eles, em termos estatísticos. Essas análises servirão de base para os modelos linguísticos que serão apresentados, de forma a inserir esta etapa no treinamento do *software* para que ele detecte o tema do que foi transcrito e adeque o contexto das palavras dentro dos trechos.

Os Quadros 15 e 16, abaixo, mostram a relação de palavras-chave processadas no *corpus* automático e no ideal. Foi utilizada uma lista de *stopwords* disponível no acervo do *GitHub*[42].

Quadro 15 - Relação das palavras-chave no automático

	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	então	35	14	pesca	13	27	dentro	9	40	certa	7
2	consumo	31	15	aula	13	28	bastante	9	41	aumento	7
3	aqui	31	16	toneladas	12	29	mundo	9	42	outro	7
4	aí	30	17	demanda	12	30	principalmente	9	43	anos	7
5	água	25	18	dados	12	31	hoje	9	44	doce	7
6	brasil	22	19	aparente	11	32	torno	8	45	importante	7
7	casa	19	20	tudo	11	33	lado	8	46	vou	7
8	pescados	18	21	produção	11	34	forma	8	47	sobre	7
9	peixe	15	22	certo	11	35	utilização	8	48	gente	7
10	oferta	14	23	pescado	10	36	fazer	8	49	conta	6
11	tá	14	24	agronegócio	10	37	potencial	8	50	consumido	6
12	grande	14	25	diferença	9	38	seguinte	8			
13	vem	13	26	coisa	9	39	fator	7			

[42] *Stopwords*: https://gist.github.com/alopes/5358189#file-stopwords-txt

Quadro 16 - Relação das palavras-chave no ideal

	Word	Frequency		Word	Frequency		Word	Frequency
1	então	39	18	demanda	12	35	torno	8
2	consumo	32	19	vem	12	36	outro	8
3	aqui	31	20	aparente	11	37	forma	8
4	água	24	21	dados	11	38	potencial	8
5	brasil	22	22	tudo	10	39	fator	7
6	tá	22	23	coisa	10	40	fazer	7
7	pescados	21	24	agronegócio	10	41	aumento	7
8	casa	20	25	diferença	9	42	lado	7
9	peixe	16	26	dentro	9	43	anos	7
10	grande	15	27	bastante	9	44	regulamentação	7
11	aula	15	28	mundo	9	45	utilização	7
12	oferta	14	29	bom	9	46	doce	7
13	toneladas	13	30	pescado	9	47	importante	7
14	certo	13	31	seguinte	9	48	vou	7
15	aquicultura	13	32	principalmente	9	49	primeira	7
16	pesca	13	33	hoje	9	50	sobre	7
17	produção	12	34	pof	8			

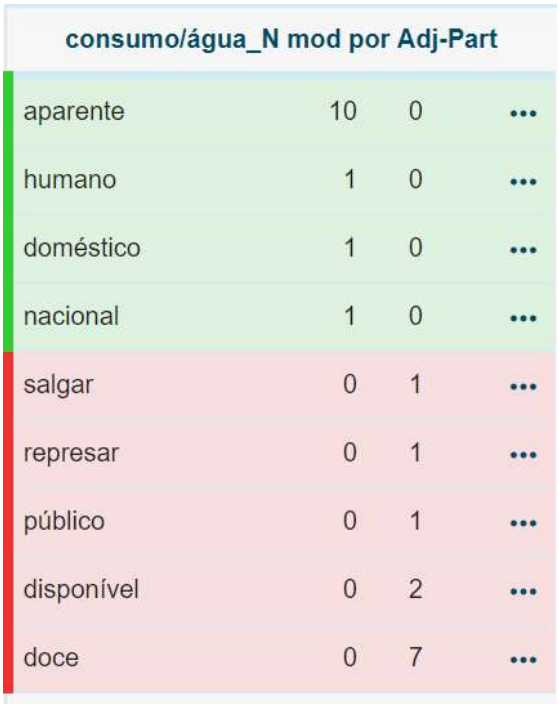
Como pode-se observar, há uma pequena diferença nos dois quadros em relação à frequência das palavras-chave processadas no automático e no ideal. A primeira palavra mais frequente nos dois *corpora* analisados foi “então”, bastante repetida durante toda a fala do professor e com uma representação um pouco maior no ideal por conta da inserção no contexto.

Algumas palavras foram repetidas em um maior volume graças às particularidades da fala do professor e suas expressões frequentes. Tais palavras, que aparecem nos dois quadros, parecem ser marcadores utilizados pelo professor, de forma a identificar itens diversos na aula. No caso do “aí”, como foi discutido na seção anterior, temos uma grande ocorrência no quadro 15 e nenhuma ocorrência no ideal. A maioria dessas repetições foram removidas no processo de adequação da transcrição por não indicarem nenhum contexto específico. Já a palavra “aqui”, presente nos dois quadros com a mesma quantidade de ocorrências, 31 vezes, pois ocorreram em relação ao contexto de outras palavras. Por fim, a palavra “tá”, ocorre com mais frequência no quadro 16, pois foi

utilizada como adequação de contexto e como inserção nos trechos em que não foram processadas essa palavra por algum tipo de erro ocorrido.

Além dessas palavras que apareceram com grande frequência nos *corpora* analisados, podemos observar outras quatro palavras relevantes e que demonstram qual o contexto da aula considerando a frequência apresentada: “consumo”, “água”, “Brasil” e “pescado”. Essas palavras-chave mostram o principal tema da aula no recorte analisado, que diz respeito exatamente ao consumo de água no Brasil para se atribuir a relação aos pescados. Ao realizar a concordância para esses trechos, pode-se observar a relação entre eles, levantando a hipótese de que, para este tema, especificamente, geralmente esses termos tendem a ocorrer crê forma conjunta. A Figura 7, abaixo, demonstra a relação entre a palavra “consumo” e a palavra “água”

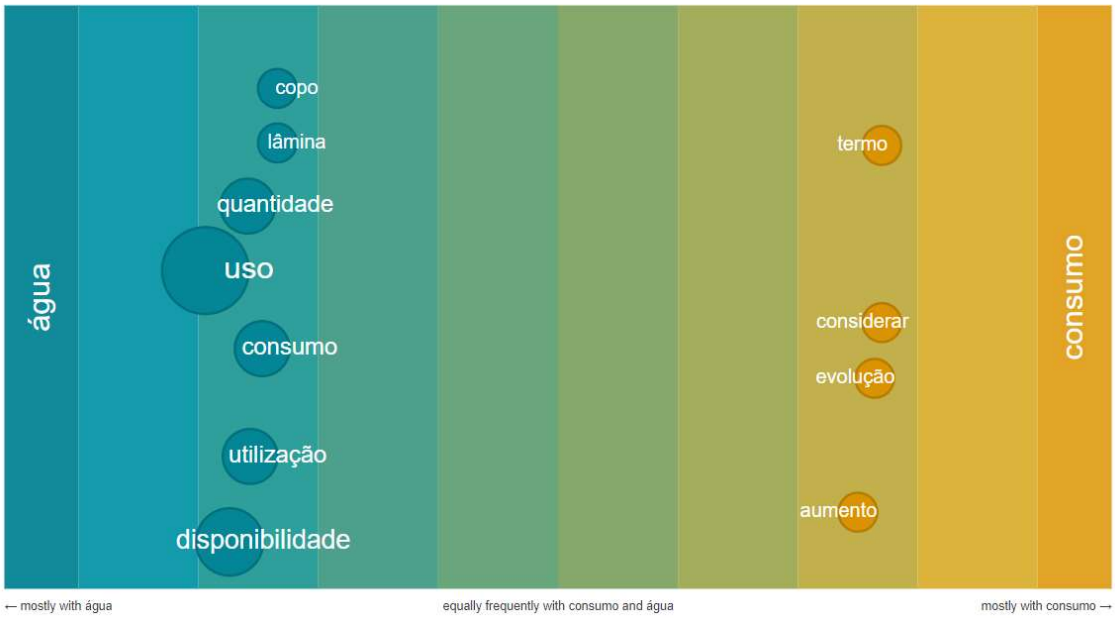
Figura 7 – Relação de “consumo” e “água”



Nota-se que a palavra “consumo” está relacionada com a palavra “aparente” e a palavra “água doce”. É importante frisar a relevância deste dado, pois através dele entende-se e pode-se inferir o contexto através da relação entre os termos, base primordial para o desenvolvimento de modelos linguísticos para o *software*. Além destas duas relações, temos outros termos relacionados que são considerados no momento de aferição

dos dados, como: “salgar”[43], “represar”, “público” e “disponível” no caso do termo “água” e “humano”, “doméstico” e “nacional” para “consumo”. Na análise do o contexto da aula, conseguimos inferir que essas palavras aparecerão, na maior parte dos casos, juntas, podendo-se levantar uma incoerência na interpretação do *software* no caso do processamento de outras palavras não correlatas. A Figura 8, assim, mostra a relação de termos entre as palavras “água” e “consumo”, expondo a correlação de outros termos entre elas.

Figura 8 - Relação de “consumo”, “água” e outros termos correlatos



Outra relação que também pode ser inferida, entre os quatro mais frequentes citados, é a de “Brasil” e “pescado”, conforme mostram as Figuras 9 e 10, abaixo.

[43] O termo correto é “salgada”, porém, como estamos utilizando a base da transcrição automática, os termos podem ser apresentados de formas incorretas.

Figura 9 - Relação de “Brasil” e “pescado”

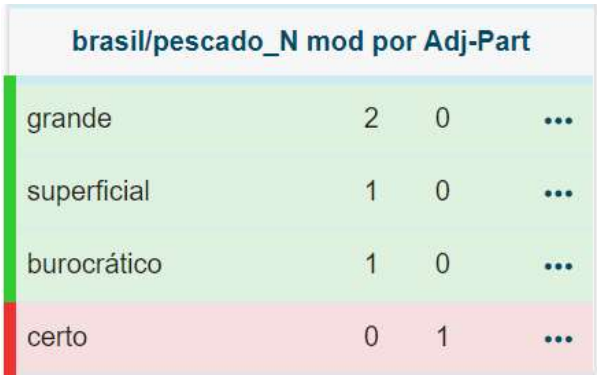


Figura 10 - Relação de “Brasil” e “pescado”

Nota-se que a relação entre as palavras nas duas figuras apresentadas varia em seus significados, denotando alguns sentidos diferentes em relação aos termos. Na Figura 9 podemos notars que, para “Brasil” as palavras relacionadas são “grande”, “superficial” e “burocrático” e para “pescado” apenas a palavra “certo”. Para “Brasil”, as palavras relacionadas estão coerentes com a explicação do professor, havendo também uma relação com as palavras da figura 10, “caso” e “balança”. O termo “Brasil” foi

constantemente associado à economia do país, processos de importação e exportação de pescados e o consumo dentro e fora de casa pelos cidadãos.

Na Figura 10 pode-se notar a relação de palavras associadas à “pescado”, como “necessidade”, “exportação”, “domiciliar”, “consumo”, “agronegócio” e “oferta”. Esses termos relacionados dão o indicativo de que “pescado” está sendo visto, nesse caso, pelo viés econômico do aproveitamento dos recursos internos para exportação ou o não aproveitamento através do consumo domiciliar. Abaixo, na Figura 11, é apresentada a relação dos dois termos: “Brasil” e “pescados”, além de outros termos correlatos.

Figura 11 - Relação de “Brasil”, “pescado” e outros termos correlatos

A próxima seção se dedica à construção dos modelos linguísticos de treinamento do *software* baseados em todas as análises quantitativas e qualitativas executadas nesta pesquisa.

4.3. Modelos de treinamento propostos para o *software*

Esta seção se dedica à demonstração dos modelos linguísticos que serão propostos para a *Skylar*. Tais modelos, discutidos nesta seção, foram embasados a partir dos resultados expostos neste capítulo.

A fim de se obter este treinamento, tanto os resultados quantitativos quanto qualitativos foram analisados. Observou-se que transcrição automática, apesar de ter um alto índice de acerto, gerava um alto número de erros, dificultando o entendimento do que o professor dizia aos alunos. Além disso, também foi apresentado também que, independente desse índice de erros e acertos, o ponto primordial era a classificação do tipo de erro atribuído ao segmento. Por isso, o estudo dos padrões dos erros das seções anteriores propiciou insumos para que melhorias fossem propostas para o treinamento da IA contida no *software*, a fim de aprimorar o processamento automático das palavras e levar a uma maior compreensão durante o processamento da língua.

Com base nisso, foram propostos 4 tipos de análises para o treinamento do *software*, sendo 1) Análise do contexto; 2) Aferição das proximidades fonéticas; 3) Contexto das palavras geradas; e 4) Aferição de ortografia das palavras.

Dados estes fatos e tendo em vista o objetivo geral deste trabalho, discute-se, a partir de agora, as nuances do modelo de treinamento propostas como meio de reduzir os erros gerados pelo *software* em questão.

Destaca-se que neste treinamento proposto o *software* será treinado com base nas palavras que ele gera, ou seja, levando-se em conta os erros identificados na geração do conteúdo automático. As próximas subseções deste capítulo detalham a proposta de treinamento elaborada neste trabalho.

4.3.1. Proposta de treinamento 1 - Análise geral do contexto

Conforme a análise apontada na seção 4.2.3, o contexto tem um papel importante na identificação de erros geração pela transcrição automática. Uma vez que o *software* processa a linguagem, mas não a compreende como tal, a análise do contexto não é levada em conta no momento entre a captação da voz e a transcrição. Este seria o primeiro passo para que o *software* passasse de processamento de linguagem natural para a compreensão de linguagem natural (ver Capítulo 2).

Nesse caso, a análise apresentada na última seção, onde podemos ver o contexto geral do *corpus* através da frequência de palavras e suas relações no contexto. Um exemplo de erro que poderia ser evitado a partir do contexto é, por exemplo, o que ocorre no segmento 403, exposto no quadro 17 abaixo.

Quadro 17 - Sequência de segmentos com erros de contexto

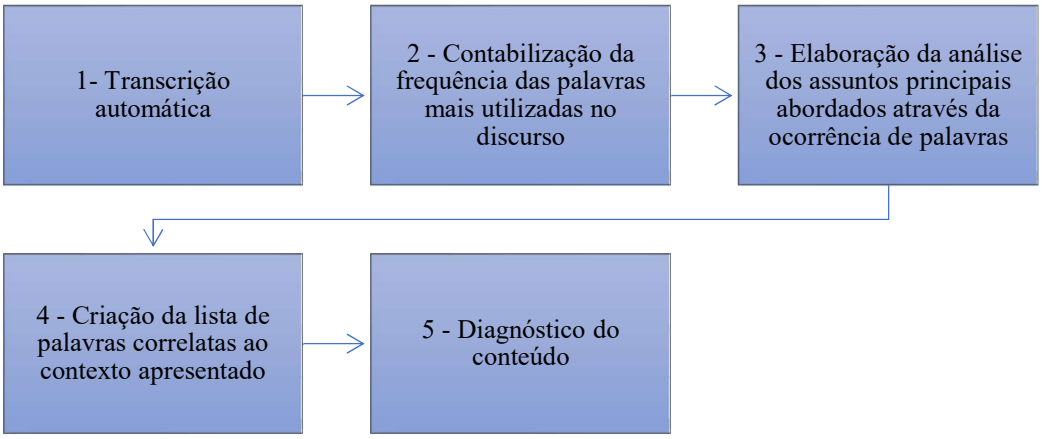
400	considerar que existem perdas e que a diferença de	considerar que existem perdas e que a diferença de
401	2002/2003 em termos de consumo dentro e fora de	2002/2003 em termos de consumo dentro e fora de
402	casa em oferta de peixe por por perdas	casa em oferta de peixe for perdas,
403	**2008/2009 eu vou ter aí uma diferença coceira nas**	**2008/2009 eu vou ter aí uma diferença, considerando as**
404	**pernas** de 265000 toneladas	perdas de 265.000 toneladas.
405	se eu considerar que olha tudo o que foi ofertado	Se eu considerar que, olha, tudo o que foi ofertado
406	foi consumido essa diferença sobe para 378000 toneladas	foi consumido, essa diferença sobe para 378.000 toneladas.

No quadro 17 há uma demonstração de todo o trecho em que ocorreu o erro do segmento 403, citado como exemplo. Hoje, o *software* não considera os outros segmentos para que sejam interpretadas as palavras que façam sentido dentro do momento de um mesmo discurso, ocasionando nesse tipo de erro. O que ocorre no 403 é um erro de similaridade fonética entre as palavras “considerando” e “coceira”, ocasionando em uma

expressão comum “coceira nas pernas”. A IA que hoje é processada no *software* busca similaridades dos conteúdos com toda a base de busca da *internet*, causando a troca de expressões específicas por expressões comuns com maior incidência na base de busca.

Portanto, a primeira proposta de treinamento levantada por essa pesquisa, foram os passos expostos no fluxograma abaixo (figura 12).

Figura 12 - Fluxo de análise de contexto



A figura 12 mostra a proposta de treinamento para o fluxo de análise de contexto de um conteúdo gerado pela *Skylar*. Após o passo 1 – a realização da transcrição do conteúdo pelo *software*, o propósito é que o *software* elabore uma aferição do contexto através de ferramentas de LC e transforme em um diagnóstico para que o *software* e os analistas saibam do que o conteúdo se trata, antes da revisão. Esse processo se apresenta de uma forma relevante para que a ferramenta possa identificar as palavras correlatas no contexto do conteúdo, buscando mitigar problemas como o apresentado nas sequências do quadro 17.

No modelo proposto, será averiguado o contexto do conteúdo e as palavras que geralmente ocorreriam neste contexto. A ideia é que se houver palavras que não se relacionem com o seu contexto ao redor ou em outros segmentos, elas sejam destacadas e enviadas para análise automática do *software* e de um analista humano treinado.

A partir dos resultados apresentados anteriormente, é possível afirmar que, ao se correlacionar o contexto geral da transcrição pode-se direcionar o *software* para a redução

de erros por falta de contexto, uma vez que as palavras fora dele serão analisadas através do contexto ao seu redor, buscando a coerência no que foi transcrito.

4.3.2. Proposta de treinamento 2 - Contexto das palavras geradas

De forma a complementar a proposta de treinamento 1, esta análise de contexto das palavras geradas é específica, isto é, analisa-se as palavras imediatamente próximas a fim de se encontrar as que não tenham qualquer relação com o contexto específico do segmento. Esta proposta é extraída do quarto procedimento da Figura 12, e é um processo fundamental para o diagnóstico gerado para a máquina.

Um exemplo é visto no segmento 250, exposto no quadro 18, em que a palavra "aquicultura" foi confundida com "apicultura".

Quadro 18 - Exemplo de palavra fora de contexto

248	que o aumento da oferta de pescador parte do	que o aumento da oferta de pescados ou parte do
249	aumento da oferta de pescados vem ocorrendo em	aumento da oferta de pescados vem ocorrendo em
250	função da **apicultura** mas sem dúvida nenhuma aqui ó	função da **aquicultura** mas, sem dúvida nenhuma aqui, olha,

Neste caso apresentado, seria inviável o uso da palavra "apicultura" através do contexto dos segmentos anteriores. Como pode-se observar nos segmentos 248 e 249, o contexto aponta para "pescado" e "oferta", que ocorre em função do cultivo de peixes, a "aquicultura" e não o cultivo de abelhas "apicultura". Além da correção também poder ser executada pela proposta de treinamento 3 de similaridade fonética, seria possível determinar e inserir uma autocorreção automática através do *software,* no caso da identificação de palavras correlatas que poderiam ocorrer no discurso.

Diante deste modelo proposto, o terceiro treinamento propõe a análise da aferição das proximidades fonéticas, procurando identificar qual o melhor modelo para mitigar este erro no *software*.

4.3.3. Proposta de treinamento 3 - Aferição das proximidades fonéticas

Os resultados obtidos na seção 4.2.1 evidenciaram que as ocorrências de proximidade fonética foram responsáveis por 26% dos erros[44] do recorte apresentado neste estudo. A partir disso, foi observado que uma análise de palavras que tenham proximidade fonética seria de grande impacto na melhoria da precisão do *software*, a fim de que este possa realizar uma escolha mais inteligente de quais palavras utilizar. O Quadro 19, abaixo, mostra um exemplo de erro por similaridade fonética.

Quadro 19 - Exemplo de sequência de erro por proximidade fonética

281	achei que a gente essa informação eu tentei fazer	A checagem dessa informação eu tentei fazer
282	através do **IBGE bebê** já tem uma pesquisa que chama	através do **IBGE. O IBGE** tem uma pesquisa que chama
283	a pesquisa de orçamento familiar que basicamente	pesquisa de orçamento familiar que basicamente

No Quadro 19 podemos observar um erro de similaridade fonética no segmento 282, causado pela repetição da palavra "IBAMA". Em um primeiro momento, justamente pelo fato de o *software* buscar correspondências do que está sendo dito na *internet*, o termo foi interpretado de forma correta, porém a sua repetição causou uma desorientação no processamento do *software*. Neste modelo proposto, busca-se a eliminação de substantivos comuns que foram confundidos com termos específicos de forma sequencial. No exemplo acima, uma regra que possibilitasse entender a relação entre "IBGE" e "pesquisa" eliminaria o erro pela similaridade fonética.

Todos os modelos propostos se complementam, de forma a entender que a elaboração da lista de palavras por contexto e de entender o próprio contexto em si, já eliminaria alguns tipos de erros causados por similaridade na fala.

Pela análise executada neste estudo, até então, pode-se identificar que os erros fonéticos causam a maioria dos problemas apresentados, sendo ele do próprio tipo 2 ou dos outros tipos. Dessa forma, pode-se constatar que o modelo de treinamento por

[44] Esta porcentagem refere-se apenas aos erros do tipo 2 isolados, não contemplando erros do tipo 2 e outros tipos combinados ou outros erros em função da similaridade fonética.

proximidade fonética é um fator que auxilia a identificação dos melhores termos do *software* de PLN analisado.

4.3.4. Proposta de treinamento 4 - Aferição de ortografia das palavras

As três classes de treinamento discutidas anteriormente permeiam a análise dos segmentos voltada para o contexto. Contudo, ainda se faz necessário que estas alterações não afetem as palavras formadas pelo treinamento. Com isso, uma última etapa do treinamento necessária para a melhoria do *software* se dá através da correção da ortografia do que foi gerado por ele. O fato de o *software* não reconhecer a junção dos caracteres como uma palavra inteira pode causar diversos tipos de problemas e erros ocasionados pela falta de letras nas palavras. Por isso, esse treinamento busca encontrar os erros ortográficos e palavras incompletas durante todo o conteúdo, inclusive das palavras oriundas das correções dos treinamentos de 1 a 3 acima.

O Quadro 20, abaixo, demonstra alguns segmentos em que as palavras não foram completadas corretamente.

Quadro 20 - Exemplo de sequência de erro com palavras incompletas

2	aula do nosso **MB** distância em Agronegócios é um	aula do nosso MBA a distância em Agronegócios.
82	**u** o porquê dessa talvez sub utilização desses	sobre o porquê dessa, talvez, subutilização desses
93	**carem** de forma ilícita é isso acabar espantando	ficarem de forma ilícita, isso acaba espantando
107	**ues**-redes dentro desses reservatórios	tanques-redes dentro desses reservatórios.
323	**ladas**	1.400.000 toneladas
330	**ansferida** para fora de casa	transferida para fora de casa?

O Quadro 20 expõe alguns exemplos de palavras processadas de forma incompleta, ocasionando em um erro de entendimento do segmento. Exceto pelo segmento 2, que apresenta uma palavra que na verdade é uma sigla da língua inglesa,

"MBA"[45], os outros segmentos apresentam palavras que não existem e estão de formas incompletas.

Neste caso, a regra sugerida para o procedimento de treinamento seria a averiguação do que foi processado para entender se, de fato, trata-se ou não uma palavra (baseado em um *corpus* com todas as palavras presentes no idioma). O *software* deve gerar o conteúdo e utilizar todos os treinamentos sugeridos acima, além de executar um tipo de comparação entre todas as palavras geradas e um *corpus* de todas as palavras do idioma de transcrição. Desta forma, a ferramenta pode identificar, por meio da comparação, se há palavras incompletas na transcrição gerada.

Se faz necessário ressaltar que os treinamentos citados neste capítulo não se dão de forma independente, ou seja, são complementares ao objetivo de reduzir os erros provenientes da geração automática de transcrição pelo *software* de reconhecimento de fala, analisado neste trabalho.

Embora os modelos de treinamento sejam provenientes do recorte analisado e limitado aos tipos de erros que foram encontrados, há margem para a elaboração de uma variação maior de análises e propostas de treinamento. Estes modelos passarão por uma adaptação, a fim de que seja compreendido o processamento computacional, o qual será aplicado, posteriormente, em forma de treinamento da IA.

As análises apresentadas nesse capítulo de resultados tiveram como objetivos estimar e avaliar a quantidade de erros ocorridas dentro do *corpus* analisado, classificar os tipos de erros e elaborar as propostas de treinamento linguísticos para a *Skylar*. Foram identificados os modelos de treinamento ideais para que o *software* aumentasse a precisão dos conteúdos gerados. Estes, por sua vez, serão aplicados na ferramenta e seus resultados serão expostos em trabalhos futuros.

---

[45] Por ser uma sigla do inglês, é possível que esse treinamento não seja efetivo para esses casos, pois geralmente o *software* confunde os idiomas na presença de um idioma diferente no segmento.

CAPÍTULO 5 – CONSIDERAÇÕES FINAIS

Esta pesquisa teve como objetivo elaborar modelos de treinamento linguístico para o software *Skylar*, através da análise dos resultados obtidos por índices de erros e contexto do conteúdo. Além da análise quantitativa, buscou-se classificar os erros de forma qualitativa, utilizando como base teórica métodos da linguística de *corpus* (BERBER SARDINHA, 2004; BIBER, 1995; BIBER; CONRAD; REPPEN, 1998; BIBER; GRAY, 2016; BIBER; REPPEN, 2015; SINCLAIR, 1991) para aferição dos dados. Para isso, foram utilizadas também métricas de linguística computacional propostas por Klakow e Peters (2002), como os índices *Word Error Rate* (WER) e Levensthein (1966), responsáveis pela contabilização da distância entre os caracteres na comparação da transcrição automática com a ideal (KLAKOW; PETERS, 2002; LEVENSTHEIN, 1966).

No *corpus* apresentado, foram apurados 406 segmentos de legendas com transcrições no português, representando 31 minutos de videoaula. O conteúdo foi compilado nos segmentos de legenda e analisados através de um software próprio para os cálculos quantitativos de forma automática. Tanto os resultados de WER, quando de LD, foram submetidos a esse processamento, a fim de se obter os números que cada um atribuía.

As perguntas que nortearam esta pesquisa foram:

1. Quais tipos de erro ocorrem durante a transcrição automática?

2. O contexto é relevante para que haja uma análise mais precisa do que precisa ser treinado na IA?

3. A estrutura do discurso do professor é representativa para que os segmentos sejam formados?

4. O formato de legendagem do *corpus* atrapalha de alguma maneira o processamento automático?

A pergunta 1 se refere à contabilização dos erros obtidos durante o processamento automático da transcrição gerada pela *Skylar.* Foi aplicado o método WER, proposto por Klakow e Peters (2002), no que concerne ao levantamento do índice de erros gerados no *corpus* analisado de forma quantitativa. Além da aplicação do método supracitado, a compilação do corpus foi baseada na fundamentação teórica apresentada no capítulo 1.

De forma a responder esta pergunta, portanto, foram levantados os principais erros e seus índices e classificações, além de se analisar as interações mais importantes entre eles a serem notadas na elaboração dos modelos de treinamento.

Para que se obtivesse a possibilidade de classificação dos erros de forma qualitativa, buscou-se, primeiramente, levantar todos os dados quantitativos em relação à performance do conteúdo automático. Foi verificado que 92% do conteúdo automático executou um desempenho considerado positivo, ou seja, com índice WER até 49%. Já para os índices de 50 a 100%, apenas 8% do corpus com esse resultado.

Embora os índices tenham apontado que a maioria do conteúdo gerado automaticamente foi positivo, levantou-se a hipótese de que somente uma análise qualitativa de cada segmento seria suficiente para entender se o software obteve mais resultados positivos que negativos, uma vez que a aplicação dessas métricas não levam em consideração o conteúdo analisado.

Interessante notar que a relação quantitativa parecia ser suficiente em um primeiro momento de reflexão. Durante a criação do software, anterior à essa pesquisa, e nas primeiras avaliações quantitativas, havia a impressão de que os resultados eram positivos, porém a experiência com a leitura das legendas sem o som do vídeo não apresentava uma compreensão que remetesse aos bons números. Tal fato motivou a análise qualitativa. Posterirormente à aplicação das métricas computacionais, no que concerne à comparação do *corpus* automático com o ideal, foram levantadas algumas questões em relações sobre as quais tipos de erros ocorrem na transcrição para se obter o real resultado da qualidade das legendas. Assim, foi elaborada uma classificação dos erros com tipos de 1 a 4.

Importante ressaltar que a elaboração de classificação dos erros ocorreu durante a interpretação dos dados, ainda na fase quantitativa. Foram levantados os erros 1) leve – alterações optativas para enquadramento nos padrões da legendagem, alteração por repetição de palavras indevidas e/ou interjeições provenientes da fala; 2) médio - palavras com proximidade fonética; e 3) grave - ausência de palavras, nomes próprios interpretados como substantivos comuns e palavras aleatórias.

Antes mesmo da classificação dos segmentos, já havia uma percepção de que os erros de similaridade fonética seriam os mais encontrados e, por conta deles, outros tipos de

erros seriam ocasionados, fato que foi demonstrado na seção de resultados. Os softwares de reconhecimento de voz buscam o aprimoramento da compreensão das palavras primeiro isoladamente, e depois por contexto (doravante discutido), justamente por já se entender ser possível que esse erro ocorra na captação dos sons das palavras.

De fato, esse tipo de erro ocorreu muitas vezes durante a análise, representando 26% de todo o conteúdo. Porém, não foi o erro com maior classificação, o tipo 1 obteve 38% e o tipo 3 com 36%. Se excluirmos o tipo 1, que pode estar relacionado ou não com similaridades fonéticas, o tipo 3, em quase sua totalidade, foi derivado de erros do tipo 2. Ou seja, os erros mais graves foram ocasionados por problemas de proximidade fonética na captação da voz.

Além desses resultados para os termos proferidos de uma forma pelo professor, e interpretado de outras formas, houve também algumas interações de hesitação na fala, que ocasionaram tipos de erros específicos por uso excessivo de jargões e expressões próprias do locutor.

A pergunta 2 refere-se à relevância do contexto para uma análise mais precisa do treinamento da IA. Buscou-se com ela observar que, ao identificar as palavras-chave e as maiores ocorrências das palavras dentro do corpus, os erros poderiam ser mitigados de forma mais efetiva.

De fato, através do processamento e levantamento dessas listas de palavras, pode-se verificar os temas principais tratados na aula. Porém, somente este dado não é suficiente para que o software comece a entender o que de fato está processando, pois ele não processa as palavras e as compreende, ou seja, não analisa e elabora uma análise de contexto. Grande parte dos problemas ocorridos poderiam ser resolvidos através da análise de contexto e de palavras correlatas, informação que foi inserida na elaboração das propostas de treinamento do software. Foi concluído, então, que a resposta da pergunta 2 tende a ser positiva, pois, através da inferência do contexto, o software não inseriria palavras aleatórias ao se isolar as palavras-chave. Na análise sugerida, as palavras mais frequentes foram “consumo”, “água”, “Brasil” e “pescado”, indicando que a aula gerava em torno destes itens (o que de fato ocorreu). Para provar estes pontos, foram realizadas análises de corpus com a ferramenta *Sketch Engine* e tais questões foram discutidas a partir das relações das palavras no contexto.

Durante toda a pesquisa, a reflexão exposta na pergunta 3 foi ponderada, buscando entender se a estrutura do discurso do locutor influenciava nos índices de erros levantados e suas implicações. Esta foi uma pergunta que, ao longo da pesquisa, foi respondida através de outros dados averiguados, mostrando que não havia uma variação de mais ou menos erros dependendo do andamento da aula. Dessa forma, apenas o contexto foi levado em consideração, e não os momentos da fala em si.

Por fim, a questão 4 levanta a reflexão acerca do formato das legendas e se este interfere no momento da comparação dos erros entre o automático e o ideal, ou se também interfere na geração dos segmentos. O tipo de erro que trata do formato de legendas é o tipo 1, o qual obteve a maior quantidade em porcentagem dos erros apresentados, 38%. Esse dado representou uma situação interessante na análise das legendas, pois vários itens que foram interpretados de forma correta pelo software tiveram alteração no trecho ideal para adequação das regras de legendagem.

Houve momentos em que os erros do tipo 1 obtiveram altos índices de WER e LD, devido a uma adequação de formato do trecho para as legendas. Assim, mostrava-se, mostrando que este formato interfere em bons resultados, em termos quantitativos. Em termos qualitativos, caso não houvesse alteração de alguns segmentos onde foram retiradas expressões utilizadas pelo professor, as quais o software processou de forma adequada, não haveria problema no entendimento das legendas.

Desta forma, a pergunta 4 foi respondida de forma a entender que o formato da legenda é importante para a compreensão dos trechos e que as alterações feitas para essa parametrização podem – ou não –, interferir no entendimento de quem consumirá as legendas.

De maneira geral, o objetivo da pesquisa, acerca da elaboração de propostas de modelos de treinamento linguístico para o software, foi cumprido na seção após a análise dos resultados. O embasamento teórico e metodológico na LC e as métricas da linguística computacional foram essenciais para que essa pesquisa fosse elaborada e que os resultados principais levantados tivessem relevância para o *corpus* analisado.

No sentido da implementação dos modelos no treinamento da IA, os quais deverão ser modificados no formato computacional para as linguagens de programação

correspondentes, a fim de que sejam viabilizados. Além deste estudo, novos estudos poderão ser elaborados a fim de que os modelos de treinamento entrem efetivamente em vigor, uma vez que algoritmos de IA desenvolvem sua própria forma de aprendizado, o que quer dizer que cabe a quem treina apresentar os parâmetros levantados por esta pesquisa.

Sendo assim, as duas linhas de pesquisa que foram utilizadas para o levantamento dos dados convergiram para que os modelos de treinamento fossem elaborados e espera-se ter levantado, nessa dissertação, a possibilidade de unir-se as linhas de pesquisa para formular um modelo híbrido de percepção dos estudos baseados em *corpus*.

Alguns fatores foram limitantes nesta pesquisa, como utilizar somente um tema para comparação da transcrição e utilizar somente um idioma e um falante. Em pesquisas futuras, busca-se ampliação destes fatores limitantes, expandindo as análises na análise de melhores modelos de treinamento para o software, depois de já ter passado pela aplicação dos modelos elaborados por esta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBATE, J. Getting Small: A Short History of the Personal Computer. **Proceedings of the IEEE**, v. 87, n. 9, p. 1695–1698, 1999.

ALLEN, J. Postediting: an integrated part of a translation software program. **Language International**, v. 13, n. April, p. 26–29, 2001.

ALMEIDA, E. M. B. Educação à distância no Brasil: diretrizes políticas, fundamentos e práticas. p. 1–6, 2002.

ALUÍSIO, S. M.; ALMEIDA, G. M. DE B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. Revista Calidoscopio. v. 4, p. 156–178, 2006.

AZIZ, W.; CASTILHO, S.; SPECIA, L. PET: a Tool for Post-editing and Assessing Machine Translation. **The Eighth International Conference on Language Resources and Evaluation**, p. 3983–3987, 2012.

BARRACHINA, S. *et al.* Statistical Approaches to Computer-Assisted Translation. **Computational.** Computational Linguistics. v. 35, n. 1, 2009.

BERBER SARDINHA, T. Linguística do Corpus: Histórico e Problemática. v. 16, p. 323–367, 2000.

BERBER SARDINHA, T. **Linguística de Corpus**. [s.l.] Manole, 2004.

BIBER, D. Representativeness in Corpus Design. v. 8, n. 4, 1993.

BIBER, D. **Dimensions of register variation: a cross-linguistic comparison**. New

York: Cambrige University Press, 1995.

BIBER, D.; CONRAD, S.; REPPEN, R. Corpus Linguistics: Investigating Language Structure and Use. p. 300, 1998.

BIBER, D.; GRAY, B. **Grammatical complexity in academic English: linguistic change in writing**. Cambrige: Cambrige University Press, 2016.

BIBER, D.; REPPEN, R. **The Cambridge handbook of English corpus linguistics**. [s.l: s.n.].

BRISCOE, T. Introduction to Linguistics for Natural Language Processing. p. 1–37, 2013.

BRITO, S. Lingüistas e computadores: que relação é essa? WORKING PAPERS EM LINGÜÍSTICA, UFSC. v. 4. 2000.

BUZATO, M. Can reading a robot derobotize a reader? **Trabalhos em Linguística Aplicada**, p. 1–9, 2010.

CINTAS, J. D.; REMAEL, A. **Audiovisual Translation: Subtitling**. [s.l.] Routledge, 2007.

CIULLA, A.; LOPES, L.; FINATTO, M. J. Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida. **Domínios da linguagem**, 2015.

CIVERA, J. et al. From Machine Translation to Computer Assisted Translation using Finite-State Models. **Proceedings of EMNLP 2004**, p. 349–356, 2004.

CRACIUNESCU, O.; SALAS, C. G.; KEEFFE, S. S. O. Machine Translation and

Computer - Assisted Translation : a New Way of Translating ? **Translation Journal**, v. 8, n. 3, p. 1–10, 2004.

DASCAL, M. **Language as a cognitive technology**. [s.l: s.n.]. Disponível em: <http://www.tau.ac.il/humanities/philos/dascal/papers/ijct-rv.htm>.

DORR, B.; SNOVER, M.; MADNAMI, N. Part 5: Machine Translation Evaluation. **Handbook of Natural Language Processing and Machine Translation**, n. 1, p. 801–894, 2010.

ESIMAJE, A. U.; HUNSTON, S. Chapter 1.1. What is corpus linguistics? In: **Studies in Corpus Linguistics**. [s.l: s.n.]. p. 8–35.

FIEDERER, R.; O'BRIEN, S. Quality and Machine Translation: A realistic objective? **The Journal of Specialised Translation**, v. 11, n. 11, p. 52–74, 2009.

GARRISON, R. Theoretical Challenges for Distance Education in the 21st Century: A Shift from Structural to Trans- actional Issues. **International Review of Research in Open and Distance Learning**, v. 1, p. 49, 2000.

GOLDBERG, D.E.; HOLLAND, J. H. Genetic Algorithms and Machine Learning. **Machine Learning**, v. 3, p. 95–99, 1988.

GOMES, M. J. Gerações de inovação tecnológica no ensino a distância. **Revista Portuguesa de Educação**, v. 43, n. 2, p. 137–156, 2003.

GUIMARÃES, J. M. DE M. Educação, globalização e educação a distância. **Revista Lusófona de Educação**, v. 9, p. 139–158, 2007.

HALLIDAY, M. A. K. Towards a theory of good translation. **Exploring translation and multilingual texts: beyond content**, p. 13–18, 2001.

HALLIDAY, M. A. K. **Computational And Quantitative Studies (Collected Works of M.A.K. Halliday Series)**. London: [s.n.].

HUTCHINS, J. Reflections on the History and Present State of Machine Translation. p. 7, 1995.

HUTCHINS, W. J. **Machine Translation: Past, Present, Future**. [s.l.] Ellis Horwood Series in Computers and their Applications, 1986.

HUTCHINS, W. J. The Georgetown-IBM experiment demonstrated in January 1954. **Proceedings of the 6th conference of the Association for Machine Translation in the Americas (AMTA 2004)**, n. January, p. 102–114, 2004.

HUTCHINS, W. J. Machine translation : a concise history. **Mechanical Translation**, v. 13, n. 1 & 2, p. 1–21, 2007.

IANNI, O. **Teorias da globalização**. 9a. ed. Rio de Janeiro: Civilização Brasileira, 2001.

JUANG, B. H.; RABINER, L. R. Automatic Speech Recognition – A Brief History of the Technology Development. p. 1–24, 2004.

KENNEDY, G. **An Introduction to Corpus Linguistics**. 1. ed. [s.l.] LONGMAN London and New York, 1998.

KENSKI, V. M. Aprendizagem mediada pela tecnologia. **Revista Diálogo Educacional**, n. 10, p. 47–56, 2003.

KHUMMONGKOL, R.; YOKOTA, M. Computer simulation of mental image processing in natural language understanding by human. **IEEE 7th International Conference on Awareness Science and Technology, iCAST 2015 - Proceedings**, p. 78–83, 2015.

KLAKOW, D.; PETERS, J. Testing the correlation of word error rate and perplexity. **Speech Communication**, v. 38, n. 1–2, p. 19–28, 2002.

LAURIAN, A. M. Machine Translation: What type of post-editing on what type of documents for what type of users. **Proceedings of the 10th Internacional Conference on Computational Linguistics and 22nd annual meeeting on Association for Computational Linguistics**, p. 236–238, 1984.

LEVENSTHEIN, V. **Binary Code Capable of Correting Deletions, Insertions, And Reversals**. [s.l: s.n.].

LIDDY, E. D. Natural language processing. **Encyclopedia of Library and Information Science**, n. 1, 2001.

LITTO, F. M. L.; FORMIGA, M. (ORG). **Educação a Distância: o estado da arte vol. 2**. 2. ed. São Paulo: Pearson, 2011.

LÜDELING, A.; KUTÖ, M. **Corpus Linguistics An International Handbook (volume 1)**. 1. ed. Berlim: [s.n.].

MCENERY, T.; HARDIE, A. Corpus linguistics: Method, theory and practice. **Corpus Linguistics: Method, Theory and Practice**, p. 1–294, 2011.

MELBY, A. Some Notes on The Proper Place of Men and Machines in Language Translation . **Machine Translation**, v. 12, n. 1–2, p. 29–34, 1997.

MORAN, J. M.; MASETTO, M. T.; BEHRENS, M. A. **Novas tecnologias e mediação pedagógica**. 10. ed. Campinas: Papirus, 2000.

MUNDAY, J. A Computer-assisted approach to the analysis of translation shifts. **Meta: Journal des traducteurs**, v. 43, n. 4, p. 542, 2012.

NAVES, S. et al. **Guia para Produções Audiovisuais Acessíveis**. [s.l: s.n.]. v. 2

NILSSON, J. et al. Parsing Formal Languages using Natural Language Parsing Techniques. **International Conference on Parsing Technologies**, n. October, p. 49–60, 2009.

OTHERO, G. DE Á. Lingüística Computacional: uma breve introdução. v. 41, n. 2, p. 341–351, 2006.

PUTRA, M. E. W.; SUPRIANA, I. Structural offline handwriting character recognition using levenshtein distance. **Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015**, v. 59, n. June 2018, p. 31–36, 2015.

RANI, N. U.; GIRIJA, P. N. Error analysis to improve the speech recognition accuracy on Telugu language. **Sadhana - Academy Proceedings in Engineering Sciences**, v. 37, n. 6, p. 747–761, 2012.

ROCHA, C. H. et al. **Anais da 1ª Jornada de Educação, Linguagem e Tecnologia**. Anais da 1a Jornada de Educação, Linguagem e Tecnologia. **Anais**...2018

SINCLAIR, J. **Corpus, Concordance, Collocation**. Hong Kong: Oxford Press University, 1991.

SKORSKA, A. et al. The CD4+AT2R+ T cell subpopulation improves post-infarction remodelling and restores cardiac function. **Journal of Cellular and Molecular Medicine**, v. 19, n. 8, p. 1975–1985, 2015.

SLOCUM, J. A Survey of Machine Translation: Its History , Current Status , and Future Prospects. **Computational Linguistics,** v. 11, n. 1, p. 1–17, 1985.

SPECIA, L.; FARZINDAR, A. Estimating machine translation post-editing effort with hter. **AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry**, p. 33–41, 2010.

Anexos

Anexo 1 - Vídeo “Agronegócio do Pescado”

Fonte: https://tinyurl.com/y2nbnxmk

Anexo 2 - Corpus de 406 segmentos da pesquisa

N. Segmento	Transcrição automática	Transcrição ideal	LD	WER	WER (%)	Tipo de erro
1	Olá pessoal boa noite como estamos aqui mais uma	Olá, pessoal, boa noite. Começamos aqui mais uma	7	0,22	22%	2
2	aula do nosso MB distância em Agronegócios é um	aula do nosso MBA a distância em Agronegócios.	8	0,44	44%	1,3
3	prazer estar aqui novamente com vocês hoje substituindo	É um prazer estar aqui novamente com vocês hoje, substituindo	6	0,25	25%	1
4	o meu colega Ricardo harbs aproveito para convidá-los	o meu colega Ricardo Harbs. Aproveito para convidá-los	1	0	0%	
5	então a participarem da nossa aula via chat	então a participarem da nossa aula via chat	0	0	0%	
6	surgirão então que envia as suas dúvidas e	sugiro então que enviem as suas dúvidas e	4	0,25	25%	2
7	perguntas durante a nossa aula via chat Vamos tornar	perguntas durante a nossa aula via chat. Vamos tornar	1	0	0%	
8	nossa aula mais dinâmica e participativo um recado	nossa aula mais dinâmica e participativa. Um recado	1	0	0%	2
9	Então antes da gente iniciaram as aulas 1º e	então antes da gente iniciar a nossa aula, primeiro e	15	0,55	55%	1,2
10	principalmente para os alunos com defesa de	principalmente para os alunos com defesa de	0	0	0%	
11	monografia prevista para dezembro deste ano eu	monografia prevista para dezembro deste ano. Eu	1	0	0%	

12	gostaria que vocês acessarem o seguinte endereço	gostaria que vocês acessassem o seguinte endereço	2	0,14	14%	2
13	www.pecege.org.br barra monografia nossa equipe de óleo só	www.pecege.esalq.usp.br/monografia. Nossa equipe de audiovisual	23	0,625	63%	1,2
14	vai colocar esse endereço para vocês depois aqui no	vai colocar esse endereço para vocês depois aqui no	0	0	0%	
15	vídeo nesse endereço estão disponíveis todos os	vídeo. Nesse endereço estão disponíveis todos os	1	0	0%	
16	manuais sobre prazos e calendários né No final do	manuais sobre prazos e calendários, né. No final do	2	0	0%	
17	ano 4 5 e 6 de dezembro será realizado o nosso	ano 4, 5 e 6 de dezembro será realizado o nosso	1	0	0%	
18	terceiro simpósio de agronegócio gestão do pecege	terceiro Simpósio de Agronegócio e Gestão do Pecege.	3	0,14	14%	1
19	então eu sugiro que você já entra nesse	Então, eu sugiro que você já entre nesse site e faça os downloads dos materiais,	42	1	100%	2,3
20	principalmente das novas normas para elaboração da	principalmente das novas normas para elaboração da	0	0	0%	
21	sua monografia bem como prazos e cronogramas também	sua monografia, bem como prazos e cronogramas. Também	2	0	0%	
22	entra em contato com seu monitor para a melhor	entre em contato com seu monitor para a melhor	1	0.11	11%	2
23	orientação hoje trataremos do assunto sobre	orientação. Hoje trataremos do assunto sobre	1	0	0%	
24	agronegócio do Pescado com professor Daniel sonora	Agronegócio do Pescado com o professor Daniel Sonoda.	4	0,29	29%	3
25	Ele é engenheiro agrônomo diretor do pecege e também	Ele é engenheiro agrônomo, diretor do Pecege e também	1	0	0%	

26	Doutor em economia aplicada pela Esalq USP desejo a	doutor em Economia Aplicada pela Esalq USP. Desejo a	1	0	0%	
27	vocês uma excelente aula	vocês uma excelente aula.	1	0	0%	
28	Boa noite a todos saudações a todos os alunos que	Boa noite a todos! Saudações a todos os alunos que	1	0,1	10%	
29	os acompanham Nesta aula de hoje é hoje vai tratar de um	nos acompanham nesta aula de hoje. Hoje a gente vai tratar de um	11	0,33	33%	1,3
30	tema polêmico é um tema interessante A então ele já	tema polêmico, um tema interessante. Então ele já	4	0,2	20%	1
31	começa a ser diferente já no título dele	começa a ser diferente no título dele:	4	0,25	25%	1
32	agronegócio do Pescado quando eu falo agronegócio do	Agronegócio do Pescado. Quando eu falo agronegócio do	1	0	0%	
33	Pescado uma vez estava ministrando as aulas lá em	pescado... uma vez estava ministrando essa aula lá em	7	0,22	22%	2
34	Manaus e a primeira pergunta que surgiu foi a seguinte é um	Manaus e a primeira pergunta que surgiu foi a seguinte: é um	1	0,08	8%	
35	agronegócio Pão Com certeza em Manaus pescado é	agronegócio? Bom, com certeza em Manaus pescado é	5	0,25	25%	2
36	muito mais oriundo da pesca do que do agronegócio mas	muito mais oriundo da pesca do que do agronegócio, mas	1	0	0%	
37	aqui na região sul e sudeste O agronegócio do	aqui na região Sul e Sudeste, o agronegócio do	1	0	0%	
38	Pescado ele vem se consolidando cada vez mais se	pescado vem se consolidando cada vez mais se	4	0,11	11%	
39	torna em um agronegócio que tem mostrado em um	torna em um agronegócio que tem mostrado em um	0	0	0%	
40	grande potencial eu vou iniciar a aula aqui com uma	grande potencial. Eu vou iniciar a aula com uma	6	0,1	10%	
41	ideia do que eu vou passar para vocês	ideia do que eu vou passar para vocês	0	0	0%	

42	durante todo o c* todo esse momento que vou ficar	durante todo esse momento que vou ficar	10	0,3	30%	1
43	junto aqui na aula tá então vou iniciar aula aqui	junto aqui na aula, tá? Então, vou iniciar aula	8	0,2	20%	1
44	com uma visão geral	com uma visão geral,	1	0	0%	
45	a motivação os pontos positivos e negativos do	a motivação, os pontos positivos e negativos do	1	0	0%	
46	agronegócio do Pescado eu gosto sempre de falar um	agronegócio do pescado. Eu gosto sempre de falar um	1	0	0%	
47	pouco de teoria e prática então é tentar unir os	pouco de teoria e prática, então é tentar unir os	1	0	0%	
48	aspectos teóricos com experiências práticas que já	aspectos teóricos com experiências práticas que	3	0,14	14%	1
49	que nós já tivemos aí em situações de em situações	nós já tivemos de situações	23	0,5	50%	1
50	aí com com empresas tá um pouquinho sobre dados e	com empresas, tá? Um pouquinho sobre dados e	9	0,3	30%	1
51	métodos principalmente aí dá onde ou Quais são as	métodos, principalmente da onde ou quais são as	5	0,22	22%	1,3
52	fontes de dados principais fontes de dados que nós	principais fontes de dados que nós	16	0,33	33%	1
53	utilizamos para fazer alisar os estudos alguns	utilizamos para realizar nossos estudos. Alguns	11	0,43	43%	2,3
54	estudos de caso para ver como é que isso funciona a	estudos de caso para ver como é que isso funciona na	1	0,09	9%	2
55	prática e dá algumas ideias uma proposta financeira	prática e dar algumas ideias, uma proposta financeira	3	0,125	13%	
56	aí para o nosso para nossa aula tá	para a nossa aula, tá?	16	0,625	63%	1
57	eu vou iniciar a aula aqui com slide do Thiago tá	Eu vou iniciar a aula com um slide do Thiago, tá?	8	0,27	27%	1

58	então levante de carne já com Tiago e a ideia é	Vocês já devem ter tido aula de carne com o Thiago. E a ideia é	30	0,82	82%	2,3
59	iniciar aí falando um pouquinho sobre a	iniciar falando um pouquinho sobre a	3	0,14	14%	1
60	disponibilidade mundial de água tá importante ter em	disponibilidade mundial de água, tá? É importante ter em	4	0,25	25%	1
61	mente o seguinte o Brasil hoje ele é um grande	mente o seguinte: o Brasil hoje é um grande	5	0,2	20%	1
62	potencial tem um grande potencial hídrico para ser	potencial, tem um grande potencial hídrico para ser	1	0	0%	
63	explorado no mundo todo a disponibilidade de água	explorado. No mundo todo, a disponibilidade de água	2	0	0%	
64	doce ela ela é muito grande no Hemisfério Sul AC da	doce é muito grande no Hemisfério Sul aqui da	11	0,27	27%	1,3
65	América mas a sua utilização ela é muito ela é muito	América, mas a sua utilização é muito	17	0,36	36%	1
66	pequena a principal a principal a principal fonte de	pequena. A principal fonte de	24	0,44	44%	1
67	forma de utilização da água é a para irrigação mas	forma de utilização da água, é para irrigação mas	3	0,1	10%	1
68	a agricultura também ela dizem	a agricultura também desempenha	10	0,4	40%	1,2,3
69	tem um papel muito importante aí na utilização	um papel muito importante na utilização	7	0,25	25%	1
70	desse recurso hídrico tá o Brasil ele é um país que	desse recurso hídrico, tá? O Brasil é um país que	6	0,18	18%	1
71	tem uma grande quantidade de água doce disponível	tem uma grande quantidade de água doce disponível.	1	0	0%	
72	essa quantidade de água doce principalmente localizada	Essa quantidade de água doce está principalmente localizada	1	0	0%	
73	nas empresas represa hidrelétrica e temos fios muito	nas represas hidrelétricas e temos rios muito	12	0,5	50%	2

74	grandes também o Brasil tem uma Costa imensa e mais	grandes também. O Brasil tem uma costa imensa, mais	3	0,1	10%	1
75	de 4 mil km Se não me engano show 8 mil km de	de 4 mil km, se não me engano mais 8 mil km de	5	0,08	8%	3
76	costa e tem mais ou menos cinco mil hectares de	costa e tem mais ou menos cinco mil hectares de	0	0	0%	
77	lâmina d'água e cinco milhões de hectares de lâmina	lâmina d'água e cinco milhões de hectares de lâmina	0	0	0%	
78	d'água para ser utilizado tá então é uma é uma é um	d'água para ser utilizado, tá? Então é um	1	0,42	42%	1
79	potencial hídrico muito grande e que até o momento	potencial hídrico muito grande e que até o momento	0	0	0%	
80	é muito pouco utilizado	é muito pouco utilizado.	1	0	0%	
81	eu queria entrar um pouco na discussão com vocês aí	Eu queria entrar um pouco na discussão com vocês aí,	1	0	0%	
82	u o porquê dessa talvez sub utilização desses	sobre o porquê dessa, talvez, subutilização desses	8	0,37	37%	3
83	recursos hídricos está atualmente no Brasil existem	recursos hídricos. Atualmente no Brasil, existem	6	0,14	14%	3
84	aí 7 agências que regulam o uso da água dentre elas o	sete agências que regulam o uso da água, dentre elas o	5	0,16	16%	1
85	mais recente o Ministério da Pesca e Agricultura	mais recente, o Ministério da Pesca e Aquicultura,	4	0,12	12%	2
86	mas a Ana que Agência Nacional de da da água você	mais a ANA que é a Agência Nacional da água. Você	12	0,45	45%	1
87	tenha você tem a Marinha Você tem os órgãos estaduais	tem a Marinha, você tem os órgãos estaduais, o IBAMA...	24	0,4	40%	2,3
88	Então são vários órgãos que regulam o uso da água	Então são vários órgãos que regulam o uso da água	0	0	0%	

89	isso acaba tornando essa utilização burocrática e	e isso acaba tornando essa utilização burocrática e	2	0,14	14%	1
90	nem sempre a regulamentação ela é rápida então como	nem sempre a regulamentação é rápida. Então como	5	0,11	11%	1
91	eu acredito que a maior parte dos empresários	eu acredito que a maior parte dos empresários	0	0	0%	
92	brasileiros não querem entrar no negócio para se	brasileiros não querem entrar em um negócio para	8	0,37	37%	3
93	carem de forma ilícita é isso acabar espantando	ficarem de forma ilícita, isso acaba espantando	5	0,37	37%	1,3
94	Muitas vezes os investidores do setor	muitas vezes os investidores do setor.	1	0	0%	
95	é importante lembrar aí também que o uso da água ela	É importante lembrar também que o uso da água	7	0,18	18%	1
96	não é só para piscicultura certo então aqui no	não é só para piscicultura, certo? Então aqui no	2	0,11	11%	
97	Brasil hoje aqui no Estado de São Paulo na região	Brasil hoje, aqui no Estado de São Paulo, na região	2	0	0%	
98	sudeste você está passando por um déficit hídrico	sudeste, você está passando por um déficit hídrico	1	0	0%	
99	muito grande e essas essas ocasiões de Déficit	muito grande e essas ocasiões de déficit	6	0,12	12%	1
100	hídrico afetam diretamente o consumo de água para	hídrico afetam diretamente o consumo de água para	0	0	0%	
101	consumo humano de água que é a principal	consumo humano de água, que é a principal	1	0	0%	
102	destinação ao destino mais nobre da da água mas	destinação, é o destino mais nobre da da água. Mas	4	0,22	22%	3
103	também ela atrapalha bastante que utiliza a água	também ela atrapalha bastante quem utiliza a água	1	0,12	12%	

104	como meio por exemplo Avicultura então é nítido aí que os	como meio, por exemplo, a avicultura. Então é nítido aí, que os	6	0,09	9%	3
105	reservatórios já estão no nível bastante baixo e	reservatórios já estão no nível bastante baixo e	0	0	0%	
106	isso vem afetando aí a produtividade de quem tem a	isso vem afetando, a produtividade de quem tem	5	0,2	20%	1
107	ues-redes dentro desses reservatórios	tanques-redes dentro desses reservatórios.	5	0,25	25%	3
108	o importante é que me motivou muito a entrar no	Uma coisa que é muito importante e que me motivou muito a entrar no	21	0,6	60%	3
109	setor isso aí já desde 98 que eu venho me dedicando a	setor, isso já desde 98 que eu venho me dedicando a	4	0,08	8%	1
110	estudar a aquicultura e pesca foi o seguinte se não	estudar a aquicultura e pesca, foi no seguinte sentido:	7	0,3	30%	3
111	seguindo sentido a pesca no mundo ela vem nos	a pesca no mundo, vem nos	21	0,33	33%	3
112	últimos ou desde a década de 90 nos últimos 20 anos	últimos, ou desde a década de 90, nos últimos 20 anos,	3	0	0%	
113	25 anos praticamente estável no mundo ela não	25 anos, praticamente estável no mundo. Ela não consegue	11	0,12	12%	1
114	consegue ou não se consegue extrair mais recursos	ou não se consegue extrair mais recursos	9	0,12	12%	1
115	pesqueiros O que é extraído atualmente por outro lado	pesqueiros do que é extraído atualmente. Por outro lado,	3	0,11	11%	2,3
116	a demanda ou oferta por pescados vem aumentando e	a demanda ou a oferta por pescados vem aumentando e	2	0,11	11%	
117	a oferta por pescados vem aumentando principalmente	a oferta por pescados vem aumentando principalmente	0	0	0%	
118	em função da aquicultura que é o que eu chamo de	em função da aquicultura, que é o que eu chamo de	1	0	0%	

119	agronegócio do Pescado então a população mundial	Agronegócio do Pescado. Então a população mundial	1	0	0%	
120	vem crescendo a demanda por pescados cresce e a	vem crescendo, a demanda por pescados cresce e a	1	0	0%	
121	pesca Mundial não consegue mais sobre esse esse	pesca mundial não consegue mais suprir esse	8	0,25	25%	2,3
122	aumento de demanda então a criação de organismos	aumento de demanda, então a criação de organismos	1	0	0%	
123	aquáticos vem sendo cada vez mais importante para	aquáticos vem sendo cada vez mais importante para	0	0	0%	
124	suprir essa necessidade	suprir essa necessidade.	1	0	0%	
125	no Brasil no Brasil o desafio é exatamente o mesmo	E no Brasil? No Brasil o desafio é exatamente o mesmo.	4	0,2	20%	2,3
126	nós aqui no Brasil temos aí um estado de de pesca	Nós aqui no Brasil temos aí um estado de pesca	3	0,09	9%	1
127	praticamente estável e o que vem ajudando aí o	praticamente estável e o que vem ajudando o	3	0,11	11%	1
128	Brasil a manter a demanda por pescados e	Brasil a manter a demanda por pescados e	0	0	0%	
129	principalmente agricultura no caso de países né hoje	principalmente a aquicultura. No caso de países, né, hoje	7	0,25	25%	2,3
130	aqui no mundo globalizado você tem muitas relações	aqui no mundo globalizado, você tem muitas relações	1	0	0%	
131	internacionais então a balança comercial do Brasil	internacionais, então a balança comercial do Brasil	1	0	0%	
132	ela é bastante importante por mais que o país não	é bastante importante. Por mais que o país não	5	0,1	10%	1
133	consiga suprir as suas necessidades de pescado como é o	consiga suprir as suas necessidades de pescado, como é o	1	0	0%	

134	caso do Brasil você tem ainda a opção de fazer a	caso do Brasil, você tem ainda a opção de fazer a	1	0	0%	
135	importação do do para suprir a sua demanda interna o	importação para suprir a sua demanda interna.	8	0,3	30%	1,3
136	ó que é importação ela está relacionada a vários	Só que a importação está relacionada a vários	6	0,33	33%	1,3
137	fatores macroeconômicos né	fatores macroeconômicos, né,	2	0	0%	
138	os fatores macroeconômicos que eu gosto de falar	os fatores macroeconômicos que eu gosto de falar	0	0	0%	
139	sempre são taxa de câmbio influencia diretamente aí a	sempre são: a taxa de câmbio influencia diretamente a	6	0,33	33%	1
140	importação e a exportação de pescados no Brasil	importação e a exportação de pescados no Brasil,	1	0	0%	
141	você tem também é a taxa de juros taxa de juros aí	você tem também a taxa de juros taxa de juros	5	0,16	16%	1
142	ela sempre influencia nos novos investimentos	que sempre influencia nos novos investimentos	3	0,16	16%	1
143	em produção	em produção	0	0	0%	
144	e aqui no Brasil você ainda tem aí as as grandes	e aqui no Brasil você ainda tem as grandes	6	0,18	18%	1
145	oportunidades que apesar de você ter muita água	oportunidades, que apesar de você ter muita água	1	0	0%	
146	disponível água doce e água salgada com potencial	disponível, água doce e água salgada, com potencial	2	0	0%	
147	de criação desses organismos Você ainda tem uma	de criação desses organismos, você ainda tem uma	1	0	0%	
148	situação aí de pouca utilização	situação de pouca utilização.	4	0,2	20%	1
149	então isso mostra que o país tem um enorme potencial	Então, isso mostra que o país tem um enorme potencial	1	0	0%	

150	aí para suprir as suas necessidades tá as suas	para suprir as suas necessidades, tá? As suas	5	0,22	22%	1
151	necessidades da sua demanda por Pescados	necessidades da sua demanda por pescados.	1	0	0%	
152	uma coisa que eu gosto de frisar sempre aí é que	Uma coisa que eu gosto de frisar sempre, é que	3	0,09	9%	1
153	é que você a gente tem aí no Brasil apesar de uma	você, a gente tem no Brasil, apesar de uma	11	0,25	25%	1,3
154	grande costuma Costa do Atlântico muito muito	grande costa, a Costa do Atlântico muito grande,	10	0,43	43%	2,3
155	é muito vasta você tem uma viscosidade aí relativamente	muito vasta, você tem uma viscosidade relativamente	6	0,22	22%	1
156	baixo Então vamos voltar rapidamente aqui a ideia	baixa. Então vamos voltar rapidamente aqui a ideia	2	0,12	12%	1,2
157	certa a ideia da aula de hoje é passar aqui uma visão	da aula de hoje, que é passar aqui uma visão	19	0,33	33%	1
158	geral fala um pouquinho da motivação pontos	geral, falar um pouquinho da motivação, pontos	3	0,14	14%	
159	positivos e pontos negativos a interação é sobre	positivos e pontos negativos, a interação é sobre	1	0	0%	
160	teoria e prática dados e métodos estudo de caso e no	teoria e prática, dados e métodos, estudo de caso e no	2	0	0%	
161	final Fallout sobre uma proposta financeira quem	final falar sobre uma proposta financeira que, quem	9	0,28	28%	2,3
162	sabe pode ser pode até ser um TCC de Algum de vocês está ao	sabe, pode ser pode até ser um TCC de algum de vocês, tá?	7	0,14	14%	3
163	meu lado e é isso que eu queria mostrar para vocês	Vamos lá! E é isso que eu queria mostrar para vocês	7	0,18	18%	3
164	que América do Sul	que a América do Sul	2	0,25	25%	1

165	ela é Ela tem um grande potencial tem uma grande	tem um grande potencial, tem uma grande	11	0,3	30%	1
166	disponibilidade de água doce por outro lado essa	disponibilidade de água doce. Por outro lado essa	1	0	0%	
167	utilização ela é muito pequena tá certo esse gráfico	utilização é muito pequena... Tá certo que esse gráfico	11	0,22	22%	1
168	aqui foi feito principalmente no caso de irrigação	aqui foi feito principalmente no caso de irrigação,	1	0	0%	
169	mas mostra que apesar de tudo nós utilizamos muito	mas mostra que apesar de tudo nós utilizamos muito	0	0	0%	
170	mal os nossos recursos	mal os nossos recursos.	1	0	0%	
171	e aqui se a ideia é olha o que que me motivou a	E aqui sim. A ideia é: "Olha o que me motivou a	9	0,33	33%	1,2
172	estudar esse setor se vocês olharem aqui ó desde a	estudar esse setor?". Se vocês olharem aqui, desde a	5	0,22	22%	1
173	década de 90 praticamente a pesca não consegue	década de 90, praticamente, a pesca não consegue	2	0,00	0%	
174	mais aumentar a sua oferta o que que vem aumentando	mais aumentar a sua oferta. O que vem aumentando	5	0,11	11%	1
175	o que vem aumentando a oferta de pescados no mundo	ou o que vem aumentando a oferta de pescados no mundo?	4	0,18	18%	3
176	basicamente é apicultura certo então aquicultura aí	Basicamente é a apicultura, certo? Então a aquicultura,	9	0,50	50%	1,2,3
177	desde a década de 90 tem sido aí a principal fonte	desde a década de 90, tem sido a principal fonte	4	0,10	10%	1
178	de aumento da oferta de pescados no mundo	de aumento da oferta de pescados no mundo.	1	0,00	0%	
179	como eu falei para vocês potencial hídrico Brasil tem aí	Como eu falei para vocês, o potencial hídrico do Brasil tem	9	0,27	27%	1

180	5.3 milhões de hectares de lâmina de água represada	5.3 milhões de hectares de lâmina de água represada	0	0,00	0%	
181	para ser utilizada represas hidrelétricas rios e	para ser utilizada represas hidrelétricas, rios e	1	0,00	0%	
182	lagos é bastante coisa é muita água doce	lagos. É bastante coisa, é muita água doce.	3	0,00	0%	
183	desde 2003 com a criação da secretaria especial da	Desde 2003 com a criação da Secretaria Especial da	0	0,00	0%	
184	aquicultura e pesca em 2009 Essa secretaria especial	Aquicultura e Pesca, e em 2009, essa secretaria especial	4	0,11	11%	1
185	virou um ministério	virou um ministério.	1	0,00	0%	
186	água ela é regulada por esse Ministério Ministério da Pesca e	A água é regulada por esse Ministério, Ministério da Pesca e	7	0,18	18%	1
187	aquicultura e também por mais seis órgãos Ministério do	Aquicultura e também por mais seis órgãos: Ministério do	1	0,11	11%	
188	meio ambiente a Marinha a Agência Nacional das Águas	Meio Ambiente, a Marinha, a Agência Nacional das Águas,	3	0,00	0%	
189	superintendência de patrimônio da União Ibama e os	a superintendência de patrimônio da União, o Ibama e os	5	0,20	20%	
190	órgãos ambientais estaduais	órgãos ambientais estaduais,	1	0,00	0%	
191	ou seja muita gente	ou seja, muita gente.	2	0,00	0%	
192	e isso causa uma certa dificuldade na hora de você	E isso causa uma certa dificuldade na hora de você	0	0,00	0%	
193	fazer a regulamentação do uso da água acaba	fazer a regulamentação do uso da água. Acaba	1	0,00	0%	
194	tornando o processo Moroso e como todo mundo já	tornando o processo moroso e, como todo mundo já	1	0,00	0%	
195	conhece o Brasil extremamente burocrático	conhece, o Brasil é extremamente burocrático.	4	0,17	17%	2

196	superficial da água dos copos de água doce podem	Bom, uma regulamentação superficial da ANA diz que 1 % da lâmina d'água dos corpos de água doce podem	53	0,63	63%	1,2,3
197	ser utilizados	ser utilizados para aquicultura, tá?	22	0,60	60%	3
198	ligação do uso da água para agricultura no Estado de São Paulo	E estão sendo criados, estão sendo feitas as regulamentações estaduais para o uso da água, ou a regulamentação para a aquicultura	95	0,90	90%	1,2,3
199	você aí a Via Rápida Paulista é um esforço da CETESB com os órgão	Aqui no Estado de São Paulo há 18 meses criou-se a Via Rápida Paulista, é um esforço da CETESB	59	0,68	68%	3
200	os órgãos estaduais eu participei de algumas	com os demais órgãos estaduais, eu participei de algumas	12	0,22	22%	1,3
201	reuniões dessa que ocorreram lá na Fiesp foi	reuniões dessas que ocorreram lá na Fiesp.	5	0,29	29%	1,2
202	amplamente discutido mais desde a sua implantação que	Foi amplamente discutido, mas desde a sua implantação, que	7	0,22	22%	2,3
203	já vai fazer mais de 18 meses a Via Rápida Paulista	já vai fazer mais de 18 meses, a Via Rápida Paulista	1	0,00	0%	
204	aprovou efetivamente poucos projetos a última	aprovou efetivamente poucos projetos. A última	1	0,00	0%	
205	notícia que eu tive foi que foi facilitado um pouco mais	notícia que eu tive foi que foi facilitado um pouco mais	0	0,00	0%	
206	esse processo para obtenção dessa licença ambiental	esse processo para obtenção dessa licença ambiental.	1	0,00	0%	
207	expectativa de todos aí principalmente daqueles que	Então a expectativa de todos, principalmente daqueles que	11	0,50	50%	1,3
208	já tô no setor é que a sua atividade seja regulamentada	já atuam no setor, é que a sua atividade seja regulamentada	5	0,09	9%	2,3
209	e ele possa trabalhar tranquilo sem aquela	e ele possa trabalhar tranquilo sem aquela	0	0,00	0%	
210	preocupação em se autuado	preocupação em se autuado.	1	0,00	0%	
211	por outro lado essa morosidade ela também atrapalha	Por outro lado, essa morosidade também atrapalha	5	0,14	14%	1

212	a entrada de novos Empreendimentos então é	a entrada de novos empreendimentos. Então é	1	0,00	0%	
213	fundamental que os órgãos estaduais e Federais e	fundamental que os órgãos estaduais, federais e	2	0,14	14%	1
214	municipais se sensibilizem na Agilidade de	municipais se sensibilizem na agilidade de	0	0,00	0%	
215	regulamentação desse processo a mente com	regulamentação desse processo. Obviamente com	6	0,40	40%	2,3
216	responsabilidade sem fazer as coisas de qualquer	responsabilidade sem fazer as coisas de qualquer	0	0,00	0%	
217	forma mas de forma responsável e rápida certo não não	forma, mas de forma responsável e rápida, certo?	10	0,38	38%	1
218	não faz sentido aguardar 1812 meses para você obter	Não faz sentido aguardar 18, 12 meses para você obter	2	0,20	20%	1,3
219	uma licença uma regulamentação para utilização de	uma licença uma regulamentação para utilização de	0	0,00	0%	
220	uma água pública está	uma água pública, tá?	4	0,25	25%	1
221	como eu falei para vocês no Brasil não é diferente	Vamos lá. Como eu falei para vocês, no Brasil não é diferente.	12	0,17	17%	3
222	a pesca estacionou aqui no patamar em torno de	A pesca estacionou aqui num patamar em torno de	2	0,11	11%	2
223	900000 toneladas por ano certo e o que tem ocorrido	900.000 toneladas por ano, certo? E o que tem ocorrido	3	0,10	10%	
224	aqui é o suplemento da da da demanda vem vem	aqui é que o suprimento da demanda vem	16	0,63	63%	1,2,3
225	ocorrendo pela oferta da aquicultura	ocorrendo pela oferta da aquicultura.	1	0,00	0%	
226	se vocês repararem nesse gráfico tem uma projeção	Se vocês repararem nesse gráfico, tem uma projeção	1	0,00	0%	
227	aqui para 2013 essa projeção para 2013	aqui para 2013. Essa projeção para 2013	1	0,00	0%	

228	Aparentemente está superestimada está mas é uma projeção	aparentemente está superestimada, tá, mas é uma projeção,	5	0,13	13%	1
229	é um estudo preliminar mas esse estudo preliminar	é um estudo preliminar, mas esse estudo preliminar	1	0,00	0%	
230	aponta o que aponta que que de fato ele está superestimado	aponta que de fato ele está superestimado.	18	0,57	57%	1
231	mas antes de tudo a pelo menos dois anos atrás eu	Mas até antes desse estudo, há pelo menos dois anos atrás eu	12	0,33	33%	2,3
232	tinha uma certa preocupação com os dados que	tinha uma certa preocupação com os dados que	0	0,00	0%	
233	estavam sendo informados em relação à produção e	estavam sendo informados em relação à produção e	0	0,00	0%	
234	ao consumo aparente de pescados no Brasil nós trabalhamos	ao consumo aparente de pescados no Brasil. Nós que trabalhamos	5	0,10	10%	1
235	com pesquisa é fundamental que as informações sejam	com pesquisa é fundamental que as informações sejam	0	0,00	0%	
236	muito criteriosos porque todas as nossas análises	muito criteriosas, porque todas as nossas análises	2	0,14	14%	2
237	e as análises econômicas que são feito em cima dessas	ou as análises econômicas que são feitas em cima dessas	4	0,20	20%	2
238	informações só dão as diretrizes ou só dão o	informações, só dão as diretrizes ou só dão o	1	0,00	0%	
239	dimensionamento correto se as informações que forem boas	dimensionamento correto se as informações forem boas.	5	0,14	14%	1
240	se as informações que podem de alguma forma	Se as informações forem de alguma forma	6	0,29	29%	2,3
241	distorcida a os resultados também eles acabam sendo	distorcidas, os resultados também acabam sendo	7	0,50	50%	1,2,3
242	não adequados tá	não adequados, tá?	2	0,33	33%	

243	uma coisa que eu gosto de chamar atenção eu acho	Uma coisa que eu gosto de chamar atenção, eu acho	1	0,00	0%	
244	muito importante é que todos estejam cientes né o	muito importante, e que todos estejam cientes, é	6	0,38	38%	3
245	Brasil ele é altamente deficitário na sua balança	o Brasil é altamente deficitário na sua balança	6	0,25	25%	1
246	comercial de pescados a gente vai mostrar para	comercial de pescados. Eu mostrei para	12	0,67	67%	3
247	vocês que a pesca no Brasil ela está estagnada e	vocês que a pesca no Brasil está estagnada e	4	0,11	11%	1
248	que o aumento da oferta de pescador parte do	que o aumento da oferta de pescados ou parte do	4	0,20	20%	2,3
249	aumento da oferta de pescados vem ocorrendo em	aumento da oferta de pescados vem ocorrendo em	0	0,00	0%	
250	função da apicultura mas sem dúvida nenhuma aqui ó	função da aquicultura mas, sem dúvida nenhuma aqui, olha,	9	0,22	22%	1,2,3
251	importação de peixe ela é muito relevante para o	a importação de peixe ela é muito relevante para o	2	0,10	10%	1
252	aumento da oferta de pescados um fator que me motiva bastante	aumento da oferta de pescados. Isso é um fator que me motiva bastante.	9	0,15	15%	3
253	Vergonha também mas o motivo é bastante porque eu	E me envergonha também. Mas me motiva bastante porque eu	13	0,60	60%	1,2,3
254	percebo que existe uma demanda reprimida por esse	percebo que existe uma demanda reprimida por esse	0	0,00	0%	
255	tipo de carne no Brasil bastante grande grande	tipo de carne no Brasil bastante grande grande.	1	0,00	0%	1
256	quanto algo em torno de quase 3 milhões de reais aula	Grande quanto? Algo em torno de quase 3 bilhões de reais ao ano, tá?	18	0,43	43%	2,3
257	está mas isso me envergonha também de certa forma	Mas isso me envergonha também de certa forma	5	0,13	13%	1,3

258	porque o Brasil é um dos maiores produtores de grãos	porque o Brasil é um dos maiores produtores de grãos	0	0,00	0%	
259	do mundo tem uma disponibilidade imensa de recursos hídricos	do mundo, tem uma disponibilidade imensa de recursos hídricos	1	0,00	0%	
260	e nós não estamos utilizando essa capacidade para	e nós não estamos utilizando essa capacidade para	0	0,00	0%	
261	produzir alimentos nós estamos importando pescados	produzir alimentos. Nós estamos importando pescados.	2	0,00	0%	
262	não faz sentido Então por um lado é um grande motivador	Não faz sentido. Então, por um lado, é um grande motivador,	4	0,00	0%	
263	demanda reprimida por uma grande oportunidade tem	mostra uma demanda reprimida, por outro uma grande oportunidade.	22	0,44	44%	1,2,3
264	coisas a serem feitas e que não estão sendo feitas	Tem coisas a serem feitas e que não estão sendo feitas.	5	0,09	9%	1
265	Então esse outro gráfico mostra a relação ao mostra	Então esse outro gráfico mostra aqui uma relação ou mostra,	10	0,30	30%	2,3
266	principalmente o consumo aparente aparente	principalmente, o consumo aparente. O que é o consumo aparente?	21	0,60	60%	3
267	8 - tudo que eu exporto então eu tenho um gráfico	O consumo aparente é a produção mais tudo que eu importo menos tudo que eu exporto.	58	0,81	81%	3
268	aqui que está mostrando o consumo aparente o consumo	Então eu tenho um gráfico aqui que está mostrando o consumo aparente. O consumo	27	0,36	36%	3
269	aparente essa linha de triângulos aqui e quem 2013	aparente é essa linha de triângulos aqui e que em 2013 é	6	0,33	33%	2,3
270	altamente influenciada pela produção e	altamente influenciada pela produção e que,	5	0,17	17%	3
271	Aparentemente está superestimada está e bastante superestimado	aparentemente, está superestimada, tá? E bastante superestimada.	7	0,29	29%	1,2
272	colchas em relação aos dados ela surgiu já há dois	Como eu falei para vocês, minha desconfiança em relação aos dados,	48	1,00	100%	1,2,3

273	anos atrás e ela surgiu em função dessa tabela aqui	surgiu já há dois anos atrás e em função dessa tabela aqui.	26	0,50	50%	1,3
274	os dados oficiais que eram fornecidos os dados	Nós temos os dados oficiais que eram fornecidos... os dados	13	0,20	20%	3
275	oficiais da oferta de pescado certo que eram	oficiais da oferta de pescado, certo, que eram	2	0,00	0%	
276	fornecidos aí pelo Ibama inicialmente depois pela	fornecidos pelo Ibama inicialmente, depois pela	4	0,17	17%	1
277	secretaria especial de Agricultura e Pesca e pelo	Secretaria Especial de Agricultura e Pesca e pelo	0	0,00	0%	2,3
278	Ministério da Pesca mais para o fim mas por outro	Ministério da Pesca mas, por fim. Mas, por outro	8	0,33	33%	2,3
279	lado quando eu calculo o consumo aparente certo eu	lado quando eu calculo o consumo aparente, certo, eu	2	0,00	0%	
280	também tenho forma de tentar chegar essa informação	também tenho uma forma de tentar checar essa informação.	6	0,22	22%	2,3
281	achei que a gente essa informação eu tentei fazer	A checagem dessa informação eu tentei fazer	13	0,57	57%	2,3
282	através do IBGE bebê já tem uma pesquisa que chama	através do IBGE. O IBGE tem uma pesquisa que chama	8	0,20	20%	
283	a pesquisa de orçamento familiar que basicamente	pesquisa de orçamento familiar que basicamente	2	0,17	17%	
284	pega tudo que um produto domicílios consome durante um ano tá	pega tudo que um domicílio consome durante um ano, tá?	11	0,30	30%	3
285	pelo consumo da Pop 100 2003 2002/2003 16 entender a	E pelo consumo da POF em 2003, 2002/2003, só para vocês entenderem	23	0,58	58%	2,3
286	porta e não é feita todo	a POF não é feita todo ano,	12	0,43	43%	2,3
287	antigamente a cada 10 anos hoje a cada seis anos em	antigamente a cada 10 anos, hoje a cada seis anos...	4	0,10	10%	3

288	2003 o consumo domiciliar de pescado sou tudo que	Então em 2003, o consumo domiciliar de pescados ou tudo que	12	0,36	36%	2,3
289	era consumido dentro de casa era alguma coisa em	era consumido dentro de casa era alguma coisa em	0	0,00	0%	
290	torno de 774000 toneladas	torno de 774.000 toneladas.	2	0,00	0%	
291	olhando os dados da evolução da produção nacional	Olhando os dados da evolução da produção nacional	0	0,00	0%	
292	ou a evolução do consumo aparente Nacional quando	ou a evolução do consumo aparente nacional quando	0	0,00	0%	
293	eu peguei os dados da POF de 2009 eu imaginei que esse	eu peguei os dados da POF de 2009 eu imaginei que esse	0	0,00	0%	
294	número aqui embaixo	número aqui embaixo	0	0,00	0%	
295	seria muito mais elevado do que é o que é	seria muito mais elevado do que	10	0,67	67%	1
296	apresentado aqui então ele praticamente se Manteve	está apresentado aqui. Então ele praticamente se manteve	6	0,13	13%	
297	constante até com uma leve queda e isso me deixou	constante até com uma leve queda e isso me deixou	0	0,00	0%	
298	muito ressabiado de certa forma porque se seus dados	muito ressabiado, de certa forma, porque se os dados	5	0,11	11%	2,3
299	apontam que existe um aumento do consumo aparente	apontam que existe um aumento do consumo aparente,	1	0,00	0%	
300	a minha expectativa era que o consumo doméstico	a minha expectativa era que o consumo doméstico ou	3	0,11	11%	
301	consumo domiciliar o consumo de pescados dentro da	consumo domiciliar, o consumo de pescados dentro da	1	0,00	0%	
302	residência aumentar as também e não foi isso que os	residência, aumentasse também e não foi isso que os	5	0,22	22%	2,3
303	dados mostraram tá então se a gente observar aqui	dados mostraram, tá? Então, se a gente observar aqui, né	7	0,20	20%	1

304	né a produção nesse período foi alguma coisa em	a produção nesse período foi alguma coisa em	3	0,13	13%	1
305	torno de 990000 toneladas o consumo	torno de 990.000 toneladas, o consumo da POF	9	0,25	25%	1
306	o consumo da pop deu 774 e a minha justificativa	deu 774 e a minha justificativa	17	0,67	67%	1
307	Inicial foi a seguinte bom	inicial foi a seguinte,	4	0,25	25%	1
308	a porta está mostrando que eu consumi dentro do	bom, a POF está mostrando que eu consumi dentro do	8	0,20	20%	
309	domicílio Então essa diferença de 774 com 990 porque	domicílio, então essa diferença de 774 com 990	8	0,13	13%	1
310	foi consumido fora domicílio Isso numa primeira uma	consumido fora do domicílio. Isso numa primeira uma	8	0,25	25%	3
311	rimeira visão me satisfez bastante mas aqui embaixo	primeira visão me satisfez bastante mas aqui embaixo	1	0,13	13%	3
312	quando eu vi que a produção pulou para um milhão e	quando eu vi que a produção pulou para um milhão e duzentas,	10	0,08	8%	1
313	200 e o consumo da pop me apontou a estabilidade	e o consumo da POF me apontou a estabilidade,	6	0,22	22%	2,3
314	comecei a ficar desconfortável Será que todo esse	comecei a ficar desconfortável. Será que todo esse	1	0,00	0%	
315	consumo excedente foi realizado fora domicílio	consumo excedente foi realizado fora do domicílio?	4	0,29	29%	3
316	como nós como eu no caso o pesquisador na área resolvi montar	Bom, então como nós, como eu no caso, sou pesquisador na área, resolvi montar	16	0,21	21%	1
317	um probleminha certo o problema que a gente se o	um probleminha, certo? Então, o problema que eu montei foi o seguinte.	27	0,50	50%	3
318	consumo aparente aparente em 2000 2013 Lembrando	Se o consumo aparente em 2002/2003, lembrando	17	0,71	71%	2,3
319	que é produção mais importação e exportação Foi algo	que é produção mais importação, menos exportação, foi algo	6	0,11	11%	3

320	em torno de 1050000 toneladas em 2002 2003 e o consumo	em torno de 1.050.000 toneladas em 2002/2003 e o consumo	3	0,20	20%	
321	da porta foi em torno de 770	da POF foi em torno de 770. Em 2008/2009	17	0,33	33%	1,2,3
322	em 2008/2009 esse consumo aparente pulo para um milhão e	esse consumo aparente pulou para	26	1,00	100%	1,2,3
323	ladas	1.400.000 toneladas	14	1,00	100%	3
324	e o meu consumo domiciliar se Manteve 770000	e o meu consumo domiciliar se manteve em 770.000.	5	0,11	11%	
325	Onde foi parar 350 mil toneladas será que tudo isso	Onde foi parar 350 mil toneladas? Será que tudo isso	1	0,10	10%	
326	foi consumido fora	foi consumido fora de casa?	9	0,40	40%	3
327	Será que o brasileiro mudou tanto hábito alimentar	Será que o brasileiro mudou tanto o hábito alimentar	2	0,11	11%	
328	assim seis anos que	assim em seis anos que	3	0,20	20%	2,3
329	a cultura de você comer peixe dentro de casa e foi	a cultura de você comer peixe dentro de casa, foi	2	0,10	10%	3
330	ansferida para fora de casa	transferida para fora de casa?	2	0,40	40%	3
331	então a minha questão era Será que o consumo fora	Bom, então a minha questão era: será que o consumo fora	6	0,18	18%	
332	de casa absorveu 630000 toneladas	de casa absorveu 630.000 toneladas?	2	0,20	20%	
333	eu comecei a fazer Conta Certa minha primeira	Eu comecei a fazer conta, certo? E minha primeira	5	0,22	22%	3
334	conta a primeira conta que eu fiz foi o seguinte	conta, a primeira conta que eu fiz, foi o seguinte.	3	0,00	0%	
335	bom então vou considerar que a oferta de pescados	Bom, então vou considerar que a oferta de pescados	1	0,00	0%	
336	ela é igual a demanda de pescados oferta de pescar	ela é igual a demanda de pescados. Então o que é a oferta de pescados?	21	0,40	40%	2,3
337	ar oferta de pescados a produção	A oferta de pescados é a produção,	4	0,29	29%	3

338	a produção	a produção,	1	0,00	0%	
339	menos ou desculpa a produção aqui mais a	menos... Ou, desculpa, a produção aqui mais a	5	0,00	0%	
340	importação menos a exportação vocês tão vendo que	importação, menos a exportação. Vocês estão vendo que	4	0,13	13%	
341	tem um F aí quem que é esse é um fator de correção	tem um F aí, que F é esse? F é um fator de correção.	9	0,29	29%	1,2,3
342	de fator de correção é esse produção peixe de peixe	Que fator de correção é esse? Quando eu falo de produção, peixe, estou falando de peixe inteiro.	46	0,53	53%	2,3
343	inteiro inteiro tem cabeça tem vísceras tem escama	Peixe inteiro tem cabeça, tem vísceras, tem escama,	8	0,13	13%	1
344	tem tudo isso	tem tudo isso.	1	0,00	0%	
345	quando eu importo peixe	Quando eu importo peixe	0	0,00	0%	
346	não faz sentido nenhum eu trazer do outro lado do	não faz sentido nenhum eu trazer do outro lado do	0	0,00	0%	
347	mundo um monte de cabeça de Cera escama Geralmente	mundo um monte de cabeça, víscera, escama. Geralmente	6	0,25	25%	2,3
348	os produtos que nos importamos ele já são	os produtos que nos importamos, eles já são	2	0,13	13%	2,3
349	minimamente beneficiados tá e a mesma coisa na	minimamente beneficiados, tá? E a mesma coisa na	2	0,13	13%	
350	exportação não faz sentido eu exportar cabeça	exportação, não faz sentido eu exportar cabeça,	2	0,00	0%	
351	vísceras escamas ou exporta alguma coisa	vísceras, escamas... eu exporto alguma coisa	6	0,33	33%	2,3
352	minimamente processados então isso tenho uma	minimamente processada. Então isso tem uma	5	0,33	33%	2,3
353	diferença certo e eu procurei jogar tudo para base	diferença, certo? E eu procurei jogar tudo para base	2	0,11	11%	

354	de peixe inteiro certo então esse fator de	de peixe inteiro, certo? Então esse fator de	2	0,13	13%	
355	correção é um fator maior do que 1 que faz o peixe	correção é um fator maior do que 1 que faz com que o peixe	8	0,14	14%	
356	já processados se seja equivalente a um peixe inteiro	já processado seja equivalente a um peixe inteiro.	5	0,25	25%	2,3
357	por outro lado a minha demanda certa a demanda que	Por outro lado, a minha demanda, certo, a demanda que	4	0,10	10%	1,2
358	eu estou considerando tudo o que eu consumi	eu estou considerando, é tudo o que eu consumi	3	0,11	11%	1
359	consumir dentro de casa mas o que eu consumi fora de	dentro de casa mais o que eu consumi fora de	10	0,20	20%	2,3
360	casa e o que eu consumi dentro de casa também tem um	casa. E o que eu consumi dentro de casa também tem um	1	0,00	0%	
361	fator de correção certo porque quando você compra	fator de correção, certo, porque quando você compra	2	0,00	0%	
362	peixe para fazer na sua casa você pode comprar peixe	peixe para fazer na sua casa você pode comprar peixe	0	0,00	0%	
363	inteiro para limpar em casa mas pode comprar peixe	inteiro para limpar em casa, mas pode comprar peixe	1	0,00	0%	
364	já minimamente processados já processado	já minimamente processado, já processado	1	0,20	20%	1
365	completamente de filé e até ele já pronto tá	completamente de filé e até ele já pronto, tá?	2	0,11	11%	
366	então minha primeira conta foi o seguinte eu queria	Então minha primeira conta foi o seguinte eu queria	0	0,00	0%	
367	achar o que que foi consumido fora de casa	achar o que foi consumido fora de casa.	5	0,13	13%	1
368	a primeira conta eu falei bom em 2000 e 2003 foram	A primeira conta eu falei, bom, em 2002/2003 foram	6	0,33	33%	2,3

369	consumidos fora de casa alguma coisa em torno de	consumidos fora de casa alguma coisa em torno de	0	0,00	0%	
370	208000 toneladas e 537000 toneladas	208.000 toneladas e 537.000 toneladas.	3	0,00	0%	
371	isso me deixou bem desconfortável falei puxa vida será	Isso me deixou bem desconfortável, falei: puxa vida, será	3	0,11	11%	
372	que o aumento foi de 158 porcento no consumo fora de casa	que o aumento foi de 158 porcento no consumo fora de casa?	1	0,08	8%	
373	então considerei de tudo consumo excedente ou toda	Então considerei que todo consumo excedente ou toda	3	0,25	25%	2,3
374	a diferença foi feita fora de casa	a diferença foi feita fora de casa.	1	0,00	0%	
375	também numa das aulas que eu tava apresentando	E também numa das aulas que eu tava apresentando	2	0,11	11%	1
376	Cidade tava meio desgostoso com essa informação um	esse dado eu tava meio desgostoso com essa informação, um	9	0,30	30%	2,3
377	aluno de pós-graduação professor uma olhada no	aluno de pós-graduação falou: "Professor, dá uma olhada no	12	0,33	33%	1
378	caderno especial da pop de 2089 que lá tem os	caderno especial da POF de 2008/2009 que lá tem os	6	0,20	20%	2,3
379	percentuais de consumo fora de casa de algumas	percentuais de consumo fora de casa de algumas	0	0,00	0%	
380	carnes e para mim o previlégio ou privilégio tinha pescados	carnes". E para meu privilégio tinha pescados.	20	0,71	71%	2,3
381	Então a partir daí eu consegui ter uma estimativa do	Então a partir daí eu consegui ter uma estimativa do	0	0,00	0%	
382	que foi consumido fora de casa também	que foi consumido fora de casa também.	1	0,00	0%	
383	e essa estimativa me mostrou que o consumo fora de	E essa estimativa me mostrou que o consumo fora de	0	0,00	0%	
384	casa em 2008/2009 foi alguma coisa em torno de	casa em 2008/2009 foi alguma coisa em torno de	0	0,00	0%	

385	142000 toneladas eu não tenho essa informação para	142.000 toneladas. Eu não tenho essa informação para	2	0,00	0%	
386	2002/2003 mas eu peguei o mesmo fator de correção e	2002/2003, mas eu peguei o mesmo fator de correção e	1	0,00	0%	
387	coloquei para 2002/2003 e deu que eu tive um	coloquei para 2002/2003 e deu que eu tive um	0	0,00	0%	
388	consumo de 114000 toneladas está	consumo de 114.000 toneladas, tá?	5	0,20	20%	1,2
389	ainda existe uma diferença	Bom, ainda existe uma diferença, tá?	10	0,33	33%	1
390	está ainda existe uma diferença	Ainda existe uma diferença.	6	0,25	25%	1
391	eu cansei dois cenários por conta dessa diferença	Eu tracei dois cenários por conta dessa diferença,	5	0,13	13%	2,3
392	eu considerei o seguinte Olha eu tenho meu consumo	eu considerei o seguinte: olha, eu tenho meu consumo	2	0,11	11%	
393	dentro de casa o meu consumo fora de casa eu tenho	dentro de casa, o meu consumo fora de casa e eu tenho	3	0,08	8%	
394	um erro que é o que esse povo tem que comer o peixe	um erro que é, esse povo quem comeu peixe	14	0,67	67%	1,2,3
395	sabe né peixe é uma carne altamente perecível então	sabe, né, peixe é uma carne altamente perecível, ela apodrece.	14	0,20	20%	3
396	tem perdas no meio desse processo o peixe estraga e	Então eu tenho perdas no meio desse processo. O peixe estraga e	13	0,25	25%	2,3
397	eu tenho que me livrar dele a ideia então é que bom	eu tenho que me livrar dele. A ideia então é que, bom,	3	0,00	0%	
398	deixa você deixa eu ser	deixa você deixa eu ser	0	0,00	0%	
399	Light aqui considerar que existem perdas se eu	mais light e aqui considerar que existem perdas. Se eu	8	0,20	20%	1
400	considerar que existem perdas e que a diferença de	considerar que existem perdas e que a diferença de	0	0,00	0%	

401	2002/2003 em termos de consumo dentro e fora de	2002/2003 em termos de consumo dentro e fora de	0	0,00	0%	
402	casa em oferta de peixe por por perdas	casa em oferta de peixe for perdas,	6	0,29	29%	1
403	2008/2009 eu vou ter aí uma diferença coceira nas	2008/2009 eu vou ter aí uma diferença, considerando as	9	0,22	22%	2,3
404	pernas de 265000 toneladas	perdas de 265.000 toneladas.	3	0,25	25%	2,3
405	se eu considerar que olha tudo o que foi ofertado	Se eu considerar que, olha, tudo o que foi ofertado	2	0,00	0%	
406	foi consumido essa diferença sobe para 378000 toneladas	foi consumido, essa diferença sobe para 378.000 toneladas.	3	0,00	0%	

Anexo 3 - Relação dos segmentos positivos (até 49% de WER) com as classificações de tipos de erro

N Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
1	Olá pessoal boa noite **como estamos** aqui mais uma	Olá, pessoal, boa noite. **Começamos** aqui mais uma	7	22%	2
2	aula do nosso **MB** distância em Agronegócios é um	aula do nosso MB**A a** distância em Agronegócios.	8	44%	1,3
3	prazer estar aqui novamente com vocês hoje substituindo	**É um** prazer estar aqui novamente com vocês hoje, substituindo	6	25%	1
4	o meu colega Ricardo harbs aproveito para convidá-los	o meu colega Ricardo Harbs. Aproveito para convidá-los	1	0%	
5	então a participarem da nossa aula via chat	então a participarem da nossa aula via chat	0	0%	
6	**surgirão** então que **envia** as suas dúvidas e	**sugiro** então que **enviem** as suas dúvidas e	4	25%	2
7	perguntas durante a nossa aula via chat Vamos tornar	perguntas durante a nossa aula via chat. Vamos tornar	1	0%	
8	nossa aula mais dinâmica e participativ**a** um recado	nossa aula mais dinâmica e participativ**a**. Um recado	1	0%	2
10	principalmente para os alunos com defesa de	principalmente para os alunos com defesa de	0	0%	
11	monografia prevista para dezembro deste ano eu	monografia prevista para dezembro deste ano. Eu	1	0%	
12	gostaria que vocês acessa**rem** o seguinte endereço	gostaria que vocês acessa**ssem** o seguinte endereço	2	14%	2
14	vai colocar esse endereço para vocês depois aqui no	vai colocar esse endereço para vocês depois aqui no	0	0%	
15	vídeo nesse endereço estão disponíveis todos os	vídeo. Nesse endereço estão disponíveis todos os	1	0%	
16	manuais sobre prazos e calendários né No final do	manuais sobre prazos e calendários, né. No final do	2	0%	
17	ano 4 5 e 6 de dezembro será realizado o nosso	ano 4, 5 e 6 de dezembro será realizado o nosso	1	0%	
18	terceiro simpósio de agronegócio gestão do pecege	terceiro Simpósio de Agronegócio **e** Gestão do Pecege.	3	14%	1

20	principalmente das novas normas para elaboração da	principalmente das novas normas para elaboração da	0	0%	
21	sua monografia bem como prazos e cronogramas também	sua monografia, bem como prazos e cronogramas. Também	2	0%	
22	entr**a** em contato com seu monitor para a melhor	entr**e** em contato com seu monitor para a melhor	1	11%	2
23	orientação hoje trataremos do assunto sobre	orientação. Hoje trataremos do assunto sobre	1	0%	
24	agronegócio do Pescado com professor Daniel **sonora**	Agronegócio do Pescado com o professor Daniel **Sonoda**.	4	29%	3
25	Ele é engenheiro agrônomo diretor do pecege e também	Ele é engenheiro agrônomo, diretor do Pecege e também	1	0%	
26	Doutor em economia aplicada pela Esalq USP desejo a	doutor em Economia Aplicada pela Esalq USP. Desejo a	1	0%	
27	vocês uma excelente aula	vocês uma excelente aula.	1	0%	
28	Boa noite a todos saudações a todos os alunos que	Boa noite a todos! Saudações a todos os alunos que	1	10%	
29	**os** acompanham Nesta aula de hoje **é** hoje vai tratar de um	**n**os acompanham nesta aula de hoje**.** Hoje **a gente** vai tratar de um	11	33%	1,3
30	tema polêmico **é** um tema interessante **A** então ele já	tema polêmico, um tema interessante. Então ele já	4	20%	1
31	começa a ser diferente **já** no título dele	começa a ser diferente no título dele:	4	25%	1
32	agronegócio do Pescado quando eu falo agronegócio do	Agronegócio do Pescado. Quando eu falo agronegócio do	1	0%	
33	Pescado uma vez estava ministrando **as aulas** lá em	pescado... uma vez estava ministrando **essa aula** lá em	7	22%	2
34	Manaus e a primeira pergunta que surgiu foi a seguinte é um	Manaus e a primeira pergunta que surgiu foi a seguinte: é um	1	8%	
35	agronegócio **Pão** Com certeza em Manaus pescado é	agronegócio? **Bom**, com certeza em Manaus pescado é	5	25%	2
36	muito mais oriundo da pesca do que do agronegócio mas	muito mais oriundo da pesca do que do agronegócio, mas	1	0%	

37	aqui na região sul e sudeste O agronegócio do	aqui na região Sul e Sudeste, o agronegócio do	1	0%	
38	Pescado ele vem se consolidando cada vez mais se	pescado vem se consolidando cada vez mais se	4	11%	
39	torna em um agronegócio que tem mostrado em um	torna em um agronegócio que tem mostrado em um	0	0%	
40	grande potencial eu vou iniciar a aula aqui com uma	grande potencial. Eu vou iniciar a aula com uma	6	10%	
41	ideia do que eu vou passar para vocês	ideia do que eu vou passar para vocês	0	0%	
42	durante todo o **c*** todo esse momento que vou ficar	durante todo esse momento que vou ficar	10	30%	1
43	junto aqui na aula tá então vou iniciar aula **aqui**	junto aqui na aula, tá? Então, vou iniciar aula	8	20%	1
44	com uma visão geral	com uma visão geral,	1	0%	
45	a motivação os pontos positivos e negativos do	a motivação, os pontos positivos e negativos do	1	0%	
46	agronegócio do Pescado eu gosto sempre de falar um	agronegócio do pescado. Eu gosto sempre de falar um	1	0%	
47	pouco de teoria e prática então é tentar unir os	pouco de teoria e prática, então é tentar unir os	1	0%	
48	aspectos teóricos com experiências práticas que **já**	aspectos teóricos com experiências práticas que	3	14%	1
50	**aí** com com empresas tá um pouquinho sobre dados e	com empresas, tá? Um pouquinho sobre dados e	9	30%	1
51	métodos principalmente **aí dá** onde ou Quais são as	métodos, principalmente da onde ou quais são as	5	22%	1,3
52	**fontes de dados** principais fontes de dados que nós	principais fontes de dados que nós	16	33%	1
53	utilizamos para **fazer alisar** os estudos alguns	utilizamos para **realizar** nossos estudos. Alguns	11	43%	2,3
54	estudos de caso para ver como é que isso funciona **a**	estudos de caso para ver como é que isso funciona **na**	1	9%	2
55	prática e dá algumas ideias uma proposta financeira	prática e dar algumas ideias, uma proposta financeira	3	13%	

57	eu vou iniciar a aula **aqui** com slide do Thiago tá	Eu vou iniciar a aula com um slide do Thiago, tá?	8	27%	1
59	iniciar **aí** falando um pouquinho sobre a	iniciar falando um pouquinho sobre a	3	14%	1
60	disponibilidade mundial de água t**á** importante ter em	disponibilidade mundial de água, tá? **É** importante ter em	4	25%	1
61	mente o seguinte o Brasil hoje **ele** é um grande	mente o seguinte: o Brasil hoje é um grande	5	20%	1
62	potencial tem um grande potencial hídrico para ser	potencial, tem um grande potencial hídrico para ser	1	0%	
63	explorado no mundo todo a disponibilidade de água	explorado. No mundo todo, a disponibilidade de água	2	0%	
64	doce **ela ela** é muito grande no Hemisfério Sul **AC** da	doce é muito grande no Hemisfério Sul aqui da	11	27%	1,3
65	América mas a sua utilização **ela** é muito **ela é muito**	América, mas a sua utilização é muito	17	36%	1
66	pequena a principal **a principal a principal** fonte de	pequena. A principal fonte,	24	44%	1
67	forma de utilização da água é **a** para irrigação mas	forma de utilização da água, é para irrigação mas	3	10%	1
68	a agricultura também **ela dizem**	a agricultura também **desempenha**	10	40%	1,2,3
69	**tem** um papel muito importante **aí** na utilização	um papel muito importante na utilização	7	25%	1
70	desse recurso hídrico tá o Brasil ele é um país que	desse recurso hídrico, tá? O Brasil é um país que	6	18%	1
71	tem uma grande quantidade de água doce disponível	tem uma grande quantidade de água doce disponível.	1	0%	
72	essa quantidade de água doce principalmente localizada	Essa quantidade de água doce está principalmente localizada	1	0%	
74	grandes também o Brasil tem uma Costa imensa e mais	grandes também. O Brasil tem uma costa imensa, mais	3	10%	1
75	de 4 mil km Se não me engano **show** 8 mil km de	de 4 mil km, se não me engano mais 8 mil km de	5	8%	3

76	costa e tem mais ou menos cinco mil hectares de	costa e tem mais ou menos cinco mil hectares de	0	0%	
77	lâmina d'água e cinco milhões de hectares de lâmina	lâmina d'água e cinco milhões de hectares de lâmina	0	0%	
78	d'água para ser utilizado tá então **é uma é uma** é um	d'água para ser utilizado, tá? Então é um	1	42%	1
79	potencial hídrico muito grande e que até o momento	potencial hídrico muito grande e que até o momento	0	0%	
80	é muito pouco utilizado	é muito pouco utilizado.	1	0%	
81	eu queria entrar um pouco na discussão com vocês aí	Eu queria entrar um pouco na discussão com vocês aí,	1	0%	
82	**u** o porquê dessa talvez sub utilização desses	sobre o porquê dessa, talvez, subutilização desses	8	37%	3
83	recursos hídricos **está** atualmente no Brasil existem	recursos hídricos. Atualmente no Brasil, existem	6	14%	3
84	**aí** 7 agências que regulam o uso da água dentre elas o	sete agências que regulam o uso da água, dentre elas o	5	16%	1
85	mais recente o Ministério da Pesca e **Agricultura**	mais recente, o Ministério da Pesca e Aquicultura,	4	12%	2
86	mas a Ana que Agência Nacional de **da da** água você	mais a ANA que é a Agência Nacional da água. Você	12	45%	1
87	**tenha** você tem a Marinha Você tem os órgãos estaduais	tem a Marinha, você tem os órgãos estaduais, **o IBAMA...**	24	40%	2,3
88	Então são vários órgãos que regulam o uso da água	Então são vários órgãos que regulam o uso da água	0	0%	
89	**i**sso acaba tornando essa utilização burocrática e	e isso acaba tornando essa utilização burocrática e	2	14%	1
90	nem sempre a regulamentação **ela** é rápida então como	nem sempre a regulamentação é rápida. Então como	5	11%	1
91	eu acredito que a maior parte dos empresários	eu acredito que a maior parte dos empresários	0	0%	
92	brasileiros não querem entrar no negócio para **se**	brasileiros não querem entrar em um negócio para	8	37%	3
93	**carem** de forma ilícita **é** isso acabar espantando	ficarem de forma ilícita, isso acaba espantando	5	37%	1,3

94	Muitas vezes os investidores do setor	muitas vezes os investidores do setor.	1	0%	
95	é importante lembrar **aí** também que o uso da água **ela**	É importante lembrar também que o uso da água	7	18%	1
96	não é só para piscicultura certo então aqui no	não é só para piscicultura, certo? Então aqui no	2	11%	
97	Brasil hoje aqui no Estado de São Paulo na região	Brasil hoje, aqui no Estado de São Paulo, na região	2	0%	
98	sudeste você está passando por um déficit hídrico	sudeste, você está passando por um déficit hídrico	1	0%	
99	muito grande e essas **essas** ocasiões de Déficit	muito grande e essas ocasiões de déficit	6	12%	1
100	hídrico afetam diretamente o consumo de água para	hídrico afetam diretamente o consumo de água para	0	0%	
101	consumo humano de água que é a principal	consumo humano de água, que é a principal	1	0%	
102	destinação **ao** destino mais nobre da da água mas	destinação, é o destino mais nobre da da água. Mas	4	22%	3
103	também ela atrapalha bastante que utiliza a água	também ela atrapalha bastante quem utiliza a água	1	12%	
104	como meio por **e**xemplo Avicultura então é nítido aí que os	como meio, por exemplo, a avicultura. Então é nítido aí, que os	6	9%	3
105	reservatórios já estão no nível bastante baixo e	reservatórios já estão no nível bastante baixo e	0	0%	
106	isso vem afetando **aí** a produtividade de quem tem **a**	isso vem afetando, a produtividade de quem tem	5	20%	1
107	**ues-redes** dentro desses reservatórios	tanques-redes dentro desses reservatórios.	5	25%	3
109	setor isso **aí** já desde 98 que eu venho me dedicando a	setor, isso já desde 98 que eu venho me dedicando a	4	8%	1
110	estudar a aquicultura e pesca foi o seguinte **se não**	estudar a aquicultura e pesca, foi no seguinte sentido:	7	30%	3
111	**seguindo sentido** a pesca no mundo ela vem nos	a pesca no mundo, vem nos	21	33%	3
112	últimos ou desde a década de 90 nos últimos 20 anos	últimos, ou desde a década de 90, nos últimos 20 anos,	3	0%	

113	25 anos praticamente estável no mundo ela não	25 anos, praticamente estável no mundo. Ela não **consegue**	11	12%	1
114	**consegue** ou não se consegue extrair mais recursos	ou não se consegue extrair mais recursos	9	12%	1
115	pesqueiros **O** que é extraído atualmente por outro lado	pesqueiros do que é extraído atualmente. Por outro lado,	3	11%	2,3
116	a demanda ou oferta por pescados vem aumentando e	a demanda ou a oferta por pescados vem aumentando e	2	11%	
117	a oferta por pescados vem aumentando principalmente	a oferta por pescados vem aumentando principalmente	0	0%	
118	em função da aquicultura que é o que eu chamo de	em função da aquicultura, que é o que eu chamo de	1	0%	
119	agronegócio do Pescado então a população mundial	Agronegócio do Pescado. Então a população mundial	1	0%	
120	vem crescendo a demanda por pescados cresce e a	vem crescendo, a demanda por pescados cresce e a	1	0%	
121	pesca Mundial não consegue mais **sobre** esse **esse**	pesca mundial não consegue mais suprir esse	8	25%	2,3
122	aumento de demanda então a criação de organismos	aumento de demanda, então a criação de organismos	1	0%	
123	aquáticos vem sendo cada vez mais importante para	aquáticos vem sendo cada vez mais importante para	0	0%	
124	suprir essa necessidade	suprir essa necessidade.	1	0%	
125	**n**o Brasil no Brasil o desafio é exatamente o mesmo	E no Brasil? No Brasil o desafio é exatamente o mesmo.	4	20%	2,3
126	nós aqui no Brasil temos aí um estado **de** de pesca	Nós aqui no Brasil temos aí um estado de pesca	3	9%	1
127	praticamente estável e o que vem ajudando **aí** o	praticamente estável e o que vem ajudando o	3	11%	1
128	Brasil a manter a demanda por pescados é	Brasil a manter a demanda por pescados é	1	0%	
129	principalmente **agricultura** no caso de países né hoje	principalmente a aquicultura. No caso de países, né, hoje	7	25%	2,3

130	aqui no mundo globalizado você tem muitas relações	aqui no mundo globalizado, você tem muitas relações	1	0%	
131	internacionais então a balança comercial do Brasil	internacionais, então a balança comercial do Brasil	1	0%	
132	**ela** é bastante importante por mais que o país não	é bastante importante. Por mais que o país não	5	10%	1
133	consiga suprir as suas necessidades de pescado como é o	consiga suprir as suas necessidades de pescado, como é o	1	0%	
134	caso do Brasil você tem ainda a opção de fazer a	caso do Brasil, você tem ainda a opção de fazer a	1	0%	
135	importação do **do** para suprir a sua demanda interna **o**	importação para suprir a sua demanda interna.	8	30%	1,3
136	**ó** que **é** importação **ela** está relacionada a vários	Só que a importação está relacionada a vários	6	33%	1,3
137	fatores macroeconômicos né	fatores macroeconômicos, né,	2	0%	
138	os fatores macroeconômicos que eu gosto de falar	os fatores macroeconômicos que eu gosto de falar	0	0%	
139	sempre são **t**axa de câmbio influencia diretamente aí a	sempre são: a taxa de câmbio influencia diretamente a	6	33%	1
140	importação e a exportação de pescados no Brasil	importação e a exportação de pescados no Brasil,	1	0%	
141	você tem também **é** a taxa de juros taxa de juros **aí**	você tem também a taxa de juros taxa de juros	5	16%	1
142	**ela** sempre influencia nos novos investimentos	**que** sempre influencia nos novos investimentos	3	16%	1
143	em produção	em produção	0	0%	
144	e aqui no Brasil você ainda tem **aí as** as grandes	e aqui no Brasil você ainda tem as grandes	6	18%	1
145	oportunidades que apesar de você ter muita água	oportunidades, que apesar de você ter muita água	1	0%	
146	disponível água doce e água salgada com potencial	disponível, água doce e água salgada, com potencial	2	0%	

147	de criação desses organismos Você ainda tem uma	de criação desses organismos, você ainda tem uma	1	0%	
148	situação **aí** de pouca utilização	situação de pouca utilização.	4	20%	1
149	então isso mostra que o país tem um enorme potencial	Então, isso mostra que o país tem um enorme potencial	1	0%	
150	**aí** para suprir as suas necessidades tá as suas	para suprir as suas necessidades, tá? As suas	5	22%	1
151	necessidades da sua demanda por Pescados	necessidades da sua demanda por pescados.	1	0%	
152	uma coisa que eu gosto de frisar sempre **aí** é que	Uma coisa que eu gosto de frisar sempre, é que	3	9%	1
153	**é que** você a gente tem **aí** no Brasil apesar de uma	você, a gente tem no Brasil, apesar de uma	11	25%	1,3
154	grande **costuma** Costa do Atlântico muito muito	grande costa, a Costa do Atlântico muito grande,	10	43%	2,3
155	**é** muito vasta você tem uma viscosidade **aí** relativamente	muito vasta, você tem uma viscosidade relativamente	6	22%	1
156	baix**o** Então vamos voltar rapidamente aqui a ideia	baixa. Então vamos voltar rapidamente aqui a ideia	2	12%	1,2
157	**certa a ideia** da aula de hoje é passar aqui uma visão	da aula de hoje, que é passar aqui uma visão	19	33%	1
158	geral fala um pouquinho da motivação pontos	geral, falar um pouquinho da motivação, pontos	3	14%	
159	positivos e pontos negativos a interação é sobre	positivos e pontos negativos, a interação é sobre	1	0%	
160	teoria e prática dados e métodos estudo de caso e no	teoria e prática, dados e métodos, estudo de caso e no	2	0%	
161	final **Fallout** sobre uma proposta financeira **quem**	final falar sobre uma proposta financeira que, quem	9	28%	2,3
162	sabe pode ser pode até ser um TCC de Algum de vocês **está ao**	sabe, pode ser pode até ser um TCC de algum de vocês, tá?	7	14%	3
163	**meu lado** e é isso que eu queria mostrar para vocês	Vamos lá! E é isso que eu queria mostrar para vocês	7	18%	3
164	que América do Sul	que **a** América do Sul	2	25%	1

165	**ela é Ela** tem um grande potencial tem uma grande	tem um grande potencial, tem uma grande	11	30%	1
166	disponibilidade de água doce por outro lado essa	disponibilidade de água doce. Por outro lado essa	1	0%	
167	utilização **ela** é muito pequena tá cert**o e**sse gráfico	utilização é muito pequena... Tá certo que esse gráfico	11	22%	1
168	aqui foi feito principalmente no caso de irrigação	aqui foi feito principalmente no caso de irrigação,	1	0%	
169	mas mostra que apesar de tudo nós utilizamos muito	mas mostra que apesar de tudo nós utilizamos muito	0	0%	
170	mal os nossos recursos	mal os nossos recursos.	1	0%	
171	e aqui **se** a ideia é olha o que **que** me motivou a	E aqui sim. A ideia é: "Olha o que me motivou a	9	33%	1,2
172	estudar esse setor se vocês olharem aqui **ó** desde a	estudar esse setor?". Se vocês olharem aqui, desde a	5	22%	1
173	década de 90 praticamente a pesca não consegue	década de 90, praticamente, a pesca não consegue	2	0%	
174	mais aumentar a sua oferta o que **que** vem aumentando	mais aumentar a sua oferta. O que vem aumentando	5	11%	1
175	**o** que vem aumentando a oferta de pescados no mundo	ou o que vem aumentando a oferta de pescados no mundo?	4	18%	3
177	desde a década de 90 tem sido **aí** a principal fonte	desde a década de 90, tem sido a principal fonte	4	10%	1
178	de aumento da oferta de pescados no mundo	de aumento da oferta de pescados no mundo.	1	0%	
179	como eu falei para você**s p**otencial hídrico Brasil tem **aí**	Como eu falei para vocês, o potencial hídrico do Brasil tem	9	27%	1
180	5.3 milhões de hectares de lâmina de água represada	5.3 milhões de hectares de lâmina de água represada	0	0%	
181	para ser utilizada em represas hidrelétricas rios e	para ser utilizada em represas hidrelétricas, rios e	1	0%	
182	lagos é bastante coisa é muita água doce	lagos. É bastante coisa, é muita água doce.	3	0%	
183	desde 2003 com a criação da secretaria especial da	Desde 2003 com a criação da Secretaria Especial da	0	0%	

184	aquicultura e pesc**a e**m 2009 Essa secretaria especial	Aquicultura e Pesca, e em 2009, essa secretaria especial	4	11%	1
185	virou um ministério	virou um ministério.	1	0%	
186	água **ela** é regulada por esse Ministério Ministério da Pesca e	A água é regulada por esse Ministério, Ministério da Pesca e	7	18%	1
187	aquicultura e também por mais seis órgãos Ministério do	Aquicultura e também por mais seis órgãos: Ministério do	1	11%	
188	meio ambiente a Marinha a Agência Nacional das Águas	Meio Ambiente, a Marinha, a Agência Nacional das Águas,	3	0%	
189	superintendência de patrimônio da União Ibama e os	a superintendência de patrimônio da União, o Ibama e os	5	20%	
190	órgãos ambientais estaduais	órgãos ambientais estaduais,	1	0%	
191	ou seja muita gente	ou seja, muita gente.	2	0%	
192	e isso causa uma certa dificuldade na hora de você	E isso causa uma certa dificuldade na hora de você	0	0%	
193	fazer a regulamentação do uso da água acaba	fazer a regulamentação do uso da água. Acaba	1	0%	
194	tornando o processo Moroso e como todo mundo já	tornando o processo moroso e, como todo mundo já	1	0%	
195	conhece o Brasil **e**xtremamente burocrático	conhece, o Brasil é extremamente burocrático.	4	17%	2
200	o**s ó**rgãos estaduais eu participei de algumas	com os demais órgãos estaduais, eu participei de algumas	12	22%	1,3
201	reuniões **dessa** que ocorreram lá na Fiesp **foi**	reuniões dessas que ocorreram lá na Fiesp.	5	29%	1,2
202	amplamente discutido **mais** desde a sua implantação que	Foi amplamente discutido, mas desde a sua implantação, que	7	22%	2,3
203	já vai fazer mais de 18 meses a Via Rápida Paulista	já vai fazer mais de 18 meses, a Via Rápida Paulista	1	0%	
204	aprovou efetivamente poucos projetos a última	aprovou efetivamente poucos projetos. A última	1	0%	
205	notícia que eu tive foi que foi facilitado um pouco mais	notícia que eu tive foi que foi facilitado um pouco mais	0	0%	

206	esse processo para obtenção dessa licença ambiental	esse processo para obtenção dessa licença ambiental.	1	0%	
208	já **tô** no setor é que a sua atividade seja regulamentada	já atuam no setor, é que a sua atividade seja regulamentada	5	9%	2,3
209	e ele possa trabalhar tranquilo sem aquela	e ele possa trabalhar tranquilo sem aquela	0	0%	
210	preocupação em ser autuado	preocupação em ser autuado.	1	0%	
211	por outro lado essa morosidade **ela** também atrapalha	Por outro lado, essa morosidade também atrapalha	5	14%	1
212	a entrada de novos Empreendimentos então é	a entrada de novos empreendimentos. Então é	1	0%	
213	fundamental que os órgãos estaduais **e** Federais e	fundamental que os órgãos estaduais, federais e	2	14%	1
214	municipais se sensibilizem na Agilidade de	municipais se sensibilizem na agilidade de	0	0%	
215	regulamentação desse processo **a mente** com	regulamentação desse processo. Obviamente com	6	40%	2,3
216	responsabilidade sem fazer as coisas de qualquer	responsabilidade sem fazer as coisas de qualquer	0	0%	
217	forma mas de forma responsável e rápida certo **não não**	forma, mas de forma responsável e rápida, certo?	10	38%	1
218	não faz sentido aguardar **1812** meses para você obter	Não faz sentido aguardar 18, 12 meses para você obter	2	20%	1,3
219	uma licença, uma regulamentação para utilização de	uma licença, uma regulamentação para utilização de	0	0%	
220	uma água pública **está**	uma água pública, tá?	4	25%	1
221	**c**omo eu falei para vocês no Brasil não é diferente	**Vamos lá**. Como eu falei para vocês, no Brasil não é diferente.	12	17%	3
222	a pesca estacionou aqui **no** patamar em torno de	A pesca estacionou aqui num patamar em torno de	2	11%	2
223	900000 toneladas por ano certo e o que tem ocorrido	900.000 toneladas por ano, certo? E o que tem ocorrido	3	10%	
225	ocorrendo pela oferta da aquicultura	ocorrendo pela oferta da aquicultura.	1	0%	

226	se vocês repararem nesse gráfico tem uma projeção	Se vocês repararem nesse gráfico, tem uma projeção	1	0%	
227	aqui para 2013 essa projeção para 2013	aqui para 2013. Essa projeção para 2013	1	0%	
228	Aparentemente está superestimada **está** mas é uma projeção	aparentemente está superestimada, tá, mas é uma projeção,	5	13%	1
229	é um estudo preliminar mas esse estudo preliminar	é um estudo preliminar, mas esse estudo preliminar	1	0%	
231	mas **antes de tudo** a pelo menos dois anos atrás eu	Mas **até antes desse estudo**, há pelo menos dois anos atrás eu	12	33%	2,3
232	tinha uma certa preocupação com os dados que	tinha uma certa preocupação com os dados que	0	0%	
233	estavam sendo informados em relação à produção e	estavam sendo informados em relação à produção e	0	0%	
234	ao consumo aparente de pescados no Brasil nó**s** **t**rabalhamos	ao consumo aparente de pescados no Brasil. Nós **que** trabalhamos	5	10%	1
235	com pesquisa é fundamental que as informações sejam	com pesquisa é fundamental que as informações sejam	0	0%	
236	muito criterios**os** porque todas as nossas análises	muito criteriosas, porque todas as nossas análises	2	14%	2
237	e as análises econômicas que são fei**to** em cima dessas	ou as análises econômicas que são fei**tas** em cima dessas	4	20%	2
238	informações só dão as diretrizes ou só dão o	informações, só dão as diretrizes ou só dão o	1	0%	
239	dimensionamento correto se as informações **que** forem boas	dimensionamento correto se as informações forem boas.	5	14%	1
240	se as informações **que** **podem** de alguma forma	Se as informações forem de alguma forma	6	29%	2,3
242	não adequados tá	não adequados, tá?	2	33%	
243	uma coisa que eu gosto de chamar atenção eu acho	Uma coisa que eu gosto de chamar atenção, eu acho	1	0%	
244	muito importante **é** que todos estejam cientes **né** **o**	muito importante, e que todos estejam cientes, é	6	38%	3

245	Brasil **ele** é altamente deficitário na sua balança	o Brasil é altamente deficitário na sua balança	6	25%	1
247	vocês que a pesca no Brasil **ela** está estagnada e	vocês que a pesca no Brasil está estagnada e	4	11%	1
248	que o aumento da oferta de **pescador** parte do	que o aumento da oferta de pescados ou parte do	4	20%	2,3
249	aumento da oferta de pescados vem ocorrendo em	aumento da oferta de pescados vem ocorrendo em	0	0%	
250	função da **apicultura** mas sem dúvida nenhuma aqui **ó**	função da aquicultura mas, sem dúvida nenhuma aqui, olha,	9	22%	1,2,3
251	**i**mportação de peixe ela é muito relevante para o	a importação de peixe é muito relevante para o	2	10%	1
252	aumento da oferta de pescados **um fator** que me motiva bastante	aumento da oferta de pescados. **Isso é um** fator que me motiva bastante.	9	15%	3
254	percebo que existe uma demanda reprimida por esse	percebo que existe uma demanda reprimida por esse	0	0%	
255	tipo de carne no Brasil bastante grande **grande**	tipo de carne no Brasil bastante grande.	1	0%	1
256	**quanto** algo em torno de quase 3 **milhões** de reais **aula**	**Grande** quanto? Algo em torno de quase 3 bilhões de reais ao ano, tá?	18	43%	2,3
257	**está** mas isso me envergonha também de certa forma	Mas isso me envergonha também de certa forma	5	13%	1,3
258	porque o Brasil é um dos maiores produtores de grãos	porque o Brasil é um dos maiores produtores de grãos	0	0%	
259	do mundo tem uma disponibilidade imensa de recursos hídricos	do mundo, tem uma disponibilidade imensa de recursos hídricos	1	0%	
260	e nós não estamos utilizando essa capacidade para	e nós não estamos utilizando essa capacidade para	0	0%	
261	produzir alimentos nós estamos importando pescados	produzir alimentos. Nós estamos importando pescados.	2	0%	
262	não faz sentido Então por um lado é um grande motivador	Não faz sentido. Então, por um lado, é um grande motivador,	4	0%	
263	**d**emanda reprimida po**r** **u**ma grande oportunidade **tem**	**mostra** uma demanda reprimida, por outro uma grande oportunidade.	22	44%	1,2,3

264	coisas a serem feitas e que não estão sendo feitas	**Tem** coisas a serem feitas e que não estão sendo feitas.	5	9%	1
265	Então esse outro gráfico mostra **a** relação **ao** mostra	Então esse outro gráfico mostra aqui uma relação ou mostra,	10	30%	2,3
268	aqui que está mostrando o consumo aparente o consumo	**Então eu tenho um gráfico** aqui que está mostrando o consumo aparente. O consumo	27	36%	3
269	aparente essa linha de triângulos aqui e **quem** 2013	aparente é essa linha de triângulos aqui e que em 2013 é	6	33%	2,3
270	altamente influenciada pela produção e	altamente influenciada pela produção e **que**,	5	17%	3
271	Aparentemente está superestimada **está** e bastante superestimad**o**	aparentemente, está superestimada, tá? E bastante superestimada.	7	29%	1,2
274	**os** dados oficiais que eram fornecidos os dados	**Nós temos** os dados oficiais que eram fornecidos... os dados	13	20%	3
275	oficiais da oferta de pescado certo que eram	oficiais da oferta de pescado, certo, que eram	2	0%	
276	fornecidos **aí** pelo Ibama inicialmente depois pela	fornecidos pelo Ibama inicialmente, depois pela	4	17%	1
277	secretaria especial de **Agricultura** e Pesca e pelo	Secretaria Especial de **Aquicultura** e Pesca e pelo	0	13%	2,3
278	Ministério da Pesca **mais para o fim** mas por outro	Ministério da Pesca mas, por fim. Mas, por outro	8	33%	2,3
279	lado quando eu calculo o consumo aparente certo eu	lado quando eu calculo o consumo aparente, certo, eu	2	0%	
280	também tenho forma de tentar **chegar** essa informação	também tenho uma forma de tentar checar essa informação.	6	22%	2,3
282	através do IBGE bebê já tem uma pesquisa que chama	através do IBGE. O IBGE tem uma pesquisa que chama	8	20%	
283	a pesquisa de orçamento familiar que basicamente	pesquisa de orçamento familiar que basicamente	2	17%	
284	pega tudo que um **produto** domicílio**s** consome durante um ano tá	pega tudo que um domicílio consome durante um ano, tá?	11	30%	3
286	**porta e** não é feita todo	a POF não é feita todo ano,	12	43%	2,3

287	antigamente a cada 10 anos hoje a cada seis anos **em**	antigamente a cada 10 anos, hoje a cada seis anos...	4	10%	3
288	2003 o consumo domiciliar de pescado **sou** tudo que	**Então em** 2003, o consumo domiciliar de pescado**s ou** tudo que	12	36%	2,3
289	era consumido dentro de casa era alguma coisa em	era consumido dentro de casa era alguma coisa em	0	0%	
290	torno de 774000 toneladas	torno de 774.000 toneladas.	2	0%	
291	olhando os dados da evolução da produção nacional	Olhando os dados da evolução da produção nacional	0	0%	
292	ou a evolução do consumo aparente Nacional quando	ou a evolução do consumo aparente nacional quando	0	0%	
293	eu peguei os dados da POF de 2009 eu imaginei que esse	eu peguei os dados da POF de 2009 eu imaginei que esse	0	0%	
294	número aqui embaixo	número aqui embaixo	0	0%	
296	apresentado aqui então ele praticamente se Manteve	está apresentado aqui. Então ele praticamente se manteve	6	13%	
297	constante até com uma leve queda e isso me deixou	constante até com uma leve queda e isso me deixou	0	0%	
298	muito ressabiado de certa forma porque **se seus** dados	muito ressabiado, de certa forma, porque **se os** dados	5	11%	2,3
299	apontam que existe um aumento do consumo aparente	apontam que existe um aumento do consumo aparente,	1	0%	
300	a minha expectativa era que o consumo doméstico	a minha expectativa era que o consumo doméstico ou	3	11%	
301	consumo domiciliar o consumo de pescados dentro da	consumo domiciliar, o consumo de pescados dentro da	1	0%	
302	residência aumen**tar as** também e não foi isso que os	residência, aumen**tasse** também e não foi isso que os	5	22%	2,3
303	dados mostraram tá então se a gente observar aqui	dados mostraram, tá? Então, se a gente observar aqui, **né**	7	20%	1
304	né a produção nesse período foi alguma coisa em	a produção nesse período foi alguma coisa em	3	13%	1

305	torno de 990000 toneladas o consum**o**	torno de 990.000 toneladas, o consumo d**a POF**	9	25%	1
307	Inicial foi a seguinte **bom**	inicial foi a seguinte,	4	25%	1
308	a **porta** está mostrando que eu consumi dentro do	bom, a **POF** está mostrando que eu consumi dentro do	8	20%	
309	domicílio Então essa diferença de 774 com 990 porque	domicílio, então essa diferença de 774 com 990	8	13%	1
310	foi consumido for**a** **d**omicílio Isso numa primeira **uma**	consumido fora do domicílio. Isso numa primeira	8	25%	3
311	**rimeira** visão me satisfez bastante mas aqui embaixo	primeira visão me satisfez bastante mas aqui embaixo	1	13%	3
312	quando eu vi que a produção pulou para um milhão e	quando eu vi que a produção pulou para um milhão **e duzentas**,	10	8%	1
313	**200** e o consumo da **pop** me apontou a estabilidade	e o consumo da POF me apontou a estabilidade,	6	22%	2,3
314	comecei a ficar desconfortável Será que todo esse	comecei a ficar desconfortável. Será que todo esse	1	0%	
315	consumo excedente foi realizado for**a** **do**micílio	consumo excedente foi realizado fora do domicílio?	4	29%	3
316	como nós como eu no caso **o** pesquisador na área resolvi montar	**Bom**, então como nós, como eu no caso, sou pesquisador na área, resolvi montar	16	21%	1
319	que é produção mais importação **e** **e**xportação Foi algo	que é produção mais importação, menos exportação, foi algo	6	11%	3
320	em torno de 1050000 toneladas em 2002 2003 e o consumo	em torno de 1.050.000 toneladas em 2002/2003 e o consumo	3	20%	
321	da **porta** foi em torno de 770	da **POF** foi em torno de 770. **Em 2008/2009**	17	33%	1,2,3
324	e o meu consumo domiciliar se Manteve 770000	e o meu consumo domiciliar se manteve em 770.000.	5	11%	
325	Onde foi parar 350 mil toneladas será que tudo isso	Onde foi parar 350 mil toneladas? Será que tudo isso	1	10%	
326	foi consumido **fora**	foi consumido fora **de casa?**	9	40%	3
327	Será que o brasileiro mudou tanto hábito alimentar	Será que o brasileiro mudou tanto o hábito alimentar	2	11%	
328	assi**m** **s**eis anos que	assim em seis anos que	3	20%	2,3

329	a cultura de você comer peixe dentro de casa **e** foi	a cultura de você comer peixe dentro de casa, foi	2	10%	3
330	**ansferida** para fora de casa	transferida para fora de casa?	2	40%	3
331	então a minha questão era Será que o consumo fora	Bom, então a minha questão era: será que o consumo fora	6	18%	
332	de casa absorveu 630000 toneladas	de casa absorveu 630.000 toneladas?	2	20%	
333	eu comecei a fazer Conta **Certa** minha primeira	Eu comecei a fazer conta, certo? E minha primeira	5	22%	3
334	conta a primeira conta que eu fiz foi o seguinte	conta, a primeira conta que eu fiz, foi o seguinte.	3	0%	
335	bom então vou considerar que a oferta de pescados	Bom, então vou considerar que a oferta de pescados	1	0%	
336	ela é igual a demanda de pescados oferta de **pescar**	ela é igual a demanda de pescados. Então o que é a oferta de pescados?	21	40%	2,3
337	**ar** oferta de pescados a produção	A oferta de pescados é a produção,	4	29%	3
338	a produção	a produção,	1	0%	
339	menos ou desculpa a produção aqui mais a	menos... Ou, desculpa, a produção aqui mais a	5	0%	
340	importação menos a exportação vocês tão vendo que	importação, menos a exportação. Vocês estão vendo que	4	13%	
341	tem um F aí **quem que é esse** é um fator de correção	tem um F aí, que F é esse? F é um fator de correção.	9	29%	1,2,3
343	inteiro **inteiro** tem cabeça tem vísceras tem escama	Peixe inteiro tem cabeça, tem vísceras, tem escama,	8	13%	1
344	tem tudo isso	tem tudo isso.	1	0%	
345	quando eu importo peixe	Quando eu importo peixe	0	0%	
346	não faz sentido nenhum eu trazer do outro lado do	não faz sentido nenhum eu trazer do outro lado do	0	0%	
347	mundo um monte de cabeça **de Cera** escama Geralmente	mundo um monte de cabeça, **víscera**, escama. Geralmente	6	25%	2,3
348	os produtos que **nos** importamos ele já são	os produtos que nós importamos, eles já são	2	13%	2,3
349	minimamente beneficiados tá e a mesma coisa na	minimamente beneficiados, tá? E a mesma coisa na	2	13%	

350	exportação não faz sentido eu exportar cabeça	exportação, não faz sentido eu exportar cabeça,	2	0%	
351	vísceras escamas **ou exporta** alguma coisa	vísceras, escamas... eu exporto alguma coisa	6	33%	2,3
352	minimamente processa**dos** então isso **tenho** uma	minimamente processada. Então isso tem uma	5	33%	2,3
353	diferença certo e eu procurei jogar tudo para base	diferença, certo? E eu procurei jogar tudo para base	2	11%	
354	de peixe inteiro certo então esse fator de	de peixe inteiro, certo? Então esse fator de	2	13%	
355	correção é um fator maior do que 1 que faz o peixe	correção é um fator maior do que 1 que faz com que o peixe	8	14%	
356	já processa**dos s**e seja equivalente a um peixe inteiro	já processado seja equivalente a um peixe inteiro.	5	25%	2,3
357	por outro lado a minha demanda **certa** a demanda que	Por outro lado, a minha demanda, certo, a demanda que	4	10%	1,2
358	eu estou considerando **tudo** o que eu consumi	eu estou considerando, é tudo o que eu consumi	3	11%	1
359	**consumir** dentro de casa **mas** o que eu consumi fora de	dentro de casa mais o que eu consumi fora de	10	20%	2,3
360	casa e o que eu consumi dentro de casa também tem um	casa. E o que eu consumi dentro de casa também tem um	1	0%	
361	fator de correção certo porque quando você compra	fator de correção, certo, porque quando você compra	2	0%	
362	peixe para fazer na sua casa você pode comprar peixe	peixe para fazer na sua casa você pode comprar peixe	0	0%	
363	inteiro para limpar em casa mas pode comprar peixe	inteiro para limpar em casa, mas pode comprar peixe	1	0%	
364	já minimamente processa**dos** já processado	já minimamente processado, já processado	1	20%	1
365	completamente de filé e até ele já pronto tá	completamente de filé e até ele já pronto, tá?	2	11%	
366	então minha primeira conta foi o seguinte eu queria	Então minha primeira conta foi o seguinte eu queria	0	0%	
367	achar o que **que** foi consumido fora de casa	achar o que foi consumido fora de casa.	5	13%	1

368	a primeira conta eu falei bom em **2000** e 2003 foram	A primeira conta eu falei, bom, em 2002/2003 foram	6	33%	2,3
369	consumidos fora de casa alguma coisa em torno de	consumidos fora de casa alguma coisa em torno de	0	0%	
370	208000 toneladas e 537000 toneladas	208.000 toneladas e 537.000 toneladas.	3	0%	
371	isso me deixou bem desconfortável falei puxa vida será	Isso me deixou bem desconfortável, falei: puxa vida, será	3	11%	
372	que o aumento foi de 158 porcento no consumo fora de casa	que o aumento foi de 158 porcento no consumo fora de casa?	1	8%	
373	então considerei **de tudo** consumo excedente ou toda	Então considerei **que todo** consumo excedente ou toda	3	25%	2,3
374	a diferença foi feita fora de casa	a diferença foi feita fora de casa.	1	0%	
375	também numa das aulas que eu tava apresentando	E também numa das aulas que eu tava apresentando	2	11%	1
376	**Cidade t**ava meio desgostoso com essa informação um	esse dado eu tava meio desgostoso com essa informação, um	9	30%	2,3
377	aluno de pós-graduaçã**o p**rofessor uma olhada no	aluno de pós-graduação **falou**: "Professor, dá uma olhada no	12	33%	1
378	caderno especial da **pop de 2089** que lá tem os	caderno especial **da POF de 2008/2009** que lá tem os	6	20%	2,3
379	percentuais de consumo fora de casa de algumas	percentuais de consumo fora de casa de algumas	0	0%	
381	Então a partir daí eu consegui ter uma estimativa do	Então a partir daí eu consegui ter uma estimativa do	0	0%	
382	que foi consumido fora de casa também	que foi consumido fora de casa também.	1	0%	
383	e essa estimativa me mostrou que o consumo fora de	E essa estimativa me mostrou que o consumo fora de	0	0%	
384	casa em 2008/2009 foi alguma coisa em torno de	casa em 2008/2009 foi alguma coisa em torno de	0	0%	
385	142000 toneladas eu não tenho essa informação para	142.000 toneladas. Eu não tenho essa informação para	2	0%	
386	2002/2003 mas eu peguei o mesmo fator de correção e	2002/2003, mas eu peguei o mesmo fator de correção e	1	0%	

387	coloquei para 2002/2003 e deu que eu tive um	coloquei para 2002/2003 e deu que eu tive um	0	0%	
388	consumo de 114000 toneladas **está**	consumo de 114.000 toneladas, tá?	5	20%	1,2
389	ainda existe uma diferença	Bom, ainda existe uma diferença, tá?	10	33%	1
390	**está** ainda existe uma diferença	Ainda existe uma diferença.	6	25%	1
391	eu **cansei** dois cenários por conta dessa diferença	Eu tracei dois cenários por conta dessa diferença,	5	13%	2,3
392	eu considerei o seguinte Olha eu tenho meu consumo	eu considerei o seguinte: olha, eu tenho meu consumo	2	11%	
393	dentro de casa o meu consumo fora de casa eu tenho	dentro de casa, o meu consumo fora de casa e eu tenho	3	8%	
395	sabe né peixe é uma carne altamente perecível então	sabe, né, peixe é uma carne altamente perecível, ela apodrece.	14	20%	3
396	**tem** perdas no meio desse processo o peixe estraga e	Então eu tenho perdas no meio desse processo. O peixe estraga e	13	25%	2,3
397	eu tenho que me livrar dele a ideia então é que bom	eu tenho que me livrar dele. A ideia então é que, bom,	3	0%	
398	deixa você deixa eu ser	deixa você, deixa eu ser	0	0%	
399	**Light** aqui considerar que existem perdas se eu	mais light e aqui considerar que existem perdas. Se eu	8	20%	1
400	considerar que existem perdas e que a diferença de	considerar que existem perdas e que a diferença de	0	0%	
401	2002/2003 em termos de consumo dentro e fora de	2002/2003 em termos de consumo dentro e fora de	0	0%	
402	casa em oferta de peixe por **por** perdas	casa em oferta de peixe por perdas,	6	29%	1
403	2008/2009 eu vou ter aí uma diferença **coceira nas**	2008/2009 eu vou ter aí uma diferença, **considerando as**	9	22%	2,3
404	**pernas** de 265000 toneladas	perdas de 265.000 toneladas.	3	25%	2,3
405	se eu considerar que olha tudo o que foi ofertado	Se eu considerar que, olha, tudo o que foi ofertado	2	0%	

406	foi consumido essa diferença sobe para 378000 toneladas	foi consumido, essa diferença sobe para 378.000 toneladas.	3	0%	

Anexo 4 - Relação dos segmentos negativos (de 50 a 100% de WER) com as classificações de tipos de erro

N. Seg.	Transcrição automática	Transcrição ideal	LD	WER (%)	Tipos de erros
9	Então antes da gente inici**aram** as aulas **1º** e	então antes da gente ini**ciar a** nossa aula, **primeiro** e	15	55%	1,2
13	www.pecege.org.br **barra** monografia nossa equipe de **óleo só**	www.pecege.esalq.usp.br/monografia. Nossa equipe de **audiovisual**	23	63%	1,2
19	então eu sugiro que você já entr**a** nesse	Então, eu sugiro que você já entre nesse **site e faça os downloads dos materiais,**	42	100%	2,3
49	**que** nós já tivemos **aí em situações** de **em** situações	nós já tivemos de situações	23	50%	1
56	**aí** para o nosso para nossa aula tá	para a nossa aula, tá?	16	63%	1
58	então **levante** de carne já co**m** Tiago e a ideia é	Vocês já devem ter tido aula de carne com o Thiago. E a ideia é	30	82%	2,3
73	nas **empresas represa** hidrelétric**a** e temos **fios** muito	nas represas hidrelétricas e temos rios muito	12	50%	2
108	o importante é que me motivou muito a entrar no	Uma coisa que é muito importante e que me motivou muito a entrar no	21	60%	3
176	basicamente é **apicultura** certo então aquicultura **aí**	Basicamente é a aquicultura, certo? Então a aquicultura,	9	50%	1,2,3
196	**superficial** da água dos **copos** de água doce podem	**Bom, uma regulamentação superficial da ANA diz que 1 % da lâmina d'água dos corpos de água doce podem**	53	63%	1,2,3
197	ser utilizados	ser utilizados para aquicultura, tá?	22	60%	3
198	ligação do uso da água para **agricultura** no Estado de São Paulo	E estão sendo criados, estão sendo feitas as regulamentações estaduais para o uso da água, ou a regulamentação para a aquicultura	95	90%	1,2,3
199	**você aí a** Via Rápida Paulista é um esforço da CETESB **com os órgão**	Aqui no Estado de São Paulo há 18 meses criou-se a Via Rápida Paulista, é um esforço da CETESB	59	68%	3
207	**e**xpectativa de todos **aí**	**Então** a axpectativa de todos, principalmente daqueles que	11	50%	1,3

	principalmente daqueles que				
224	aqui é o **suplemento da da** da demanda vem **vem**	aqui é que o suprimento da demanda vem	16	63%	1,2,3
230	aponta **o que** aponta que **que** de fato ele está superestimado	aponta que de fato ele está superestimado.	18	57%	1
241	distorcid**a a** os resultados também **eles** acabam sendo	distorcidas, os resultados também acabam sendo	7	50%	1,2,3
246	comercial de pescados **a gente vai mostrar para**	comercial de pescados. **Eu mostrei para**	12	67%	3
253	**Vergonha** também mas **o motivo é** bastante porque eu	**E me envergonha** também. Mas me motiva bastante porque eu	13	60%	1,2,3
266	principalmente o consumo aparente **aparente**	principalmente, o consumo aparente. O que é o consumo aparente?	21	60%	3
267	**8 -** tudo que eu exporto então **eu tenho um gráfico**	**O consumo aparente é a produção mais tudo que eu importo menos** tudo que eu exporto.	58	81%	3
272	colchas em relação aos dados ela surgiu já há dois	Como eu falei para vocês, minha desconfiança em relação aos dados,	48	100%	1,2,3
273	anos atrás e ela surgiu em função dessa tabela aqui	surgiu já há dois anos atrás e em função dessa tabela aqui.	26	50%	1,3
281	**achei que a gente essa informação** eu tentei fazer	A checagem dessa informação eu tentei fazer	13	57%	2,3
285	pelo consumo da **Pop 100** 2003 2002/2003 **16** entender a	**E** pelo consumo da **POF** em 2003, 2002/2003, só para vocês **entenderem**	23	58%	2,3
295	seria muito mais elevado do que **é o que é**	seria muito mais elevado do que	10	67%	1
306	o consumo **da pop** deu 774 e a minha justificativa	deu 774 e a minha justificativa	17	67%	1
317	um probleminha certo o **problema que a gente se o**	um probleminha, certo? Então, o problema que eu montei foi o seguinte.	27	50%	3
318	consumo aparente aparente em 2000 2013 Lembrando	**Se o** consumo aparente em 2002/2003, lembrando	17	71%	2,3

322	**em 2008/2009** esse consumo aparente **pulo para um milhão e**	esse consumo aparente pulou para	26	100%	1,2,3
323	**ladas**	**1.400.000 toneladas**	14	100%	3
342	**de** fator de correção é esse **produção peixe de peixe**	Que fator de correção é esse? Quando eu falo de produção, peixe, estou falando de peixe inteiro.	46	53%	2,3
380	carnes e para **mim o** previlégio ou privilégio tinha pescados	carnes". E para **meu** previlégio tinha pescados.	20	71%	2,3
394	um erro que **é o que esse povo tem que comer o peixe**	um erro que é, esse povo que come peixe	14	67%	1,2,3