

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LINGUÍSTICA

**INTELIGÊNCIA SEMIÓTICA E TRIPADVISOR:
DEEP LEARNING E QUANTIFICAÇÕES SUBJETIVAS NA ANÁLISE
DOS DISCURSOS DE SANÇÃO DE RESTAURANTE**

TULIO FERREIRA LEITE DA SILVA

Dissertação apresentada ao Programa de Pós-Graduação em Semiótica e Linguística Geral do Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo para a obtenção do título de Mestre em Linguística.

Orientador: Prof. Dr. Waldir Bevidas

VERSÃO CORRIGIDA

SÃO PAULO

2021

**INTELIGÊNCIA SEMIÓTICA E TRIPADVISOR
DEEP LEARNING E QUANTIFICAÇÕES SUBJETIVAS NA ANÁLISE
DOS DISCURSOS DE SANÇÃO DE RESTAURANTE**

TULIO FERREIRA LEITE DA SILVA

VERSÃO CORRIGIDA

SÃO PAULO

2021

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

S586i Silva, Tulio Ferreira Leite da
Inteligência semiótica e TripAdvisor: deep
learning e quantificações subjetivas na análise dos
discursos de sanção de restaurante / Tulio Ferreira
Leite da Silva; orientador Waldir Bevidas - São
Paulo, 2021.
126 f.

Dissertação (Mestrado)- Faculdade de Filosofia,
Letras e Ciências Humanas da Universidade de São
Paulo. Departamento de Linguística. Área de
concentração: Semiótica e Lingüística Geral.

1. Processamento de Linguagem Natural. 2.
Semiótica. 3. Análise de Sentimento. 4. Aprendizado
de Máquina. 5. Gastronomia. I. Bevidas, Waldir,
orient. II. Título.

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Ciência e Concordância do orientador

Nome do aluno: Tulio Ferreira Leite da Silva

Data da defesa: 11/03/2021

Nome do Prof. orientador: Waldir Bevidas

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento e publicação no **Portal Digital de Teses da USP**.

São Paulo, 07/05/2021



Prof. Dr. Waldir Bevidas

DA SILVA, Tulio Ferreira Leite. **Inteligência semiótica e TripAdvisor: deep Learning e quantificações subjetivas na análise dos discursos de sanção de restaurante.** Dissertação apresentada à Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo para a obtenção do título de doutor em Semiótica e Linguística Geral.

Aprovação em: 11/03/2021.

Banca examinadora

Prof. Dr.: Waldir Beividas

Instituição: FFLCH - USP

Presidente da Banca

Profa. Dra.: Ana Cristina Fricke Matte

Instituição: UFMG - Externo

Julgamento: Aprovação

Prof. Dr.: Luiz Sergio de Souza

Instituição: FATEC - Externo

Julgamento: Aprovação

Prof. Dr.: Felipe Iszlaji de Albuquerque

Instituição: FIAP - Externo

Julgamento: Aprovação

A Deus! Há deus? Adeus...

AGRADECIMENTOS

Aos meus pais, pela estrutura que sempre me permitiu voar seguindo minha biruta interna. A meus irmãos, de sangue e de consideração (Lívia e Alcides), pela paciente companhia na viagem. À Denise, por evitar que uma pane elétrica resultasse em perda total. Ao meu orientador, Prof. Dr. Waldir Beividas, pela total liberdade na confecção do plano de voo. Ao Prof. Dr. Ivã Carlos Lopes e à Dra. Paula Martins, pelo paraquedas, quando a queda qualificativa parecia iminente (e novamente à Dra. Paula, pela generosidade junto a e-mails longuíssimos). Ao Prof. Dr. Luiz Tatit que (sem saber) forneceu um combustível inesperado.

Ao Instituto de Matemática e Estatística (IME) e ao USP CodeLab, principalmente aos meus colegas de InterHack (Evandro Nakayama, Gabriel Brandão e Rogério Fernandes) por me aceitarem em um grupo de programação quando eu sequer sabia o que era a linha de comando. Nesse sentido, obrigado também ao Prof. Dr. Marcos Lopes, que tantas dúvidas me tirou na formatação desta ideia.

À Escola Politécnica (POLI-USP) sou grato por toda a formação complementar. Principalmente ao Laboratório de Arquitetura e Redes de Computadores (LARC), que me concedeu bolsa nos cursos EAD de *Big Data* e de *Machine Learning*. Ao PECE (Programa de Educação Continuada da Escola Politécnica), muito obrigado pela bolsa (de dois anos) na especialização em "*Big Data: Inteligência na Gestão dos Dados*" e por me unir a colegas e professores com quem aprendo sempre.

À equipe técnica da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) e à secretaria do Departamento de Linguística, representados por Juliana Oliveira, Érica Flávia de Lima e Robson Dantas Vieira: obrigado pela paciência com todas as dúvidas e pelos direcionamentos sempre precisos e preciosos.

Ao GES-USP, minha profunda gratidão pelo ensino da política e da alteridade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES): muito obrigado pela confiança, figurativizada em forma de bolsa, sem a qual este trabalho nunca teria saído do mundo das ideias.

À Alexandra Elbakyan, que tornou nossas pesquisas mais fáceis.

Quanto mais apurada é a medida, mais indireta ela é. A ciência do solitário é qualitativa. A ciência socializada é quantitativa. A dualidade Universo e Espírito, quando examinada no âmbito de um esforço de conhecimento pessoal, aparece como a dualidade do fenômeno mal preparado e da sensação não retificada. A mesma dualidade fundamental, quando examinada no âmbito de um esforço de conhecimento científico, aparece como a dualidade do aparelho e da teoria, dualidade já não em oposição mas em recíprocas.

Gaston Bachelard

RESUMO

Uma realidade em que internet móvel 5G estará disponível em qualquer lugar do planeta bate à porta. Se a tecnologia anterior (4G) foi diretamente responsável pela possibilidade da revolução causada por Waze, Uber e iFood, ainda não somos capazes de prever o cenário dos próximos anos. A única certeza que temos: haverá muito texto para processar. Nossa contribuição com esta pesquisa, portanto, é a de fornecer atualização a uma proposta dos anos 1980 de utilização da semiótica como ferramenta de representação do conhecimento a ser utilizado por inteligências artificiais no processamento de linguagem natural (PLN). Assim, partimos das contribuições heurísticas fornecidas pela Escola de Paris e de novidades propostas pelo francês Claude Zilberberg, criador da hipótese tensiva. Por meio desse aparato teórico, apresentamos uma análise da estrutura das avaliações de restaurante na plataforma TripAdvisor e propomos um modelo de classificação para colaborar como pré-processamento para a análise de sentimentos. A partir de um corpus manualmente anotado com essa nova proposta, utilizamos elementos do estado da arte em inteligência artificial para PLN (deep learning) e implementamos um sistema automático de classificação. O modelo, desenvolvido para fazer as vezes de prova de conceito, alcança mais de 90% de acurácia. Entretanto, não são poucas as evidências de que esteja havendo sobreajuste (overfitting). Novas pesquisas estão em andamento para sanar este ponto, ainda assim, todo o aparato teórico não é afetado pelos problemas técnicos. Mais do que isso, ainda debatemos neste trabalho formas de vencer tais dificuldades por meio da classificação de mais corpus.

Palavras-chave: Processamento de Linguagem Natural, Semiótica, Análise de Sentimento, Aprendizado de Máquina, Gastronomia.

ABSTRACT

Worldwide availability of 5G mobile internet is fast becoming a reality. If the previous technology (4G) was directly responsible for enabling the revolution of Waze, Uber and iFood, we still can't predict the next few years. The only certainty we have is that there'll be a whole lot of text to process. Our contribution with this research, therefore, is to update the 1980's proposal of using semiotics as a tool of knowledge representation to aid artificial intelligence in natural language processing (NLP). Thus, we start with the heuristical contributions provided by the School of Paris, and the new proposals from Claude Zilberberg, creator of the tensive hypothesis. By means of this theoretical apparatus, we present a structural analysis of restaurant reviews on the TripAdvisor platform, and we propose a classification model to collaborate with the pre-processing of sentiment analysis. From a corpus manually notated with this new proposal, we use state of the art elements in NLP artificial intelligence (deep learning) to implement an automatic classification system. The model, developed to stand as a proof of concept, reaches over 90% accuracy. However, there is no shortage of evidence of overfitting, which new ongoing research is working to remedy. Still, the theoretical apparatus remains unaffected by these technical issues. Beyond that, we debate in this paper manners of overcoming such difficulties through the classification of more texts.

Keywords: Natural Language Processing, Semiotics, Sentiment Analysis, Machine Learning, Gastronomy.

SUMÁRIO

0. INTRODUÇÃO	11
0.1. Motivação	12
0.2. Objetivo	14
0.3. Justificativa	14
0.4. Contribuição	15
0.5. Material e método	15
0.6. Organização do trabalho	15
1. HISTORIOGRAFIA DA INTERFACE ENTRE SEMIÓTICA E COMPUTAÇÃO	17
1.1. Algirdas Greimas e Ciências da Computação	20
1.2. Pierre Maranda e a Análise Computacional do Mito	25
1.3. Madeleine Arnold e Inteligência Artificial.....	33
2. MODELO PRESSUPOSTO DA AVALIAÇÃO DE RESTAURANTE	42
2.1. Objeto “avaliação de restaurante no TripAdvisor”	43
2.2. Contexto situacional e estrutura actancial do enunciado	53
2.3. Tipologia das sanções de restaurante	60
3. MANIFESTAÇÃO: NOVOS ACERCAMENTOS À SUPERFÍCIE DO TEXTO	70
3.1. Análise de sentimento.....	70
3.2. Quantificações subjetivas	81
3.2.1. Exemplos de classificação	85
3.3. Prova de conceito com rede neural.....	99
4. CONCLUSÃO	108
4.1. Contribuições do trabalho.....	109
4.2. Perspectivas futuras	110
REFERÊNCIAS BIBLIOGRÁFICA	111
A.POC: RELATÓRIO DE TREINAMENTO DE INTELIGÊNCIA ARTIFICIAL	115

0. INTRODUÇÃO

A semiótica da Escola de Paris, proposta por Algirdas Julien Greimas a partir de Propp e das análises narrativas do conto maravilhoso, bem como dos trabalhos do antropólogo Claude Lévi-Strauss, ainda hoje apresenta grande valor heurístico para as análises do discurso.

"Um projeto de ciências", era como Greimas costumava se referir à semiótica que se desenvolvia debaixo de seus olhos, sempre preocupados com a imanência da linguagem. Conhecedor da tendência centrífuga das Ciências Humanas, o pesquisador teve o cuidado de unificar metodologias e metalinguagem. Exemplo máximo desta característica é o *Dicionário de Semiótica* escrito junto a Joseph Courtés. A obra, que se diz tributária de Louis Trolle Hjelmslev, apresenta grande organicidade e homogeneidade entre todos os seus termos. Por conta disso, tornou-se porto-seguro para o pesquisador que busca iniciar seus estudos do discurso e da linguagem.

Verdadeiramente visionário, já nos primeiros anos da década de 1980, Greimas organizou um seminário em 10 sessões, de duas horas cada, para debater as interfaces entre semiótica e inteligência artificial. Do lado das tecnologias, Jean Zeitoun, criador de inúmeros centros de informática e tecnologia para o Estado francês, ficou responsável por apresentar os avanços da IA.

Inúmeros trabalhos foram publicados após esse encontro. Entre eles encontram-se artigos publicados por Madeleine Arnold, que propõe a utilização da semiótica na representação dos conhecimentos a serem utilizados em inteligência artificial.

Entretanto, as pesquisas realizadas nos anos 1980 parecem ainda não terem sido adaptadas à nova realidade dos anos 2020. Imensas quantidades de dados verbais produzidos nas redes sociais digitais (que deixariam Lévi-Strauss assombrado), crescente facilidade na programação computacional (introduzida principalmente pela linguagem Python), computação na nuvem (que reduz a necessidade de o pesquisador dispor de máquinas muito poderosas), são apenas as principais mudanças a que os trabalhos de Madeleine Arnold precisam ser confrontados.

Somado a isso, a teoria semiótica vem sofrendo um grande e positivo incremento metodológico e heurístico por meio dos trabalhos do francês Claude Zilberberg, que desenvolveu a hipótese tensiva. Por meio desses trabalhos, uma nova metalinguagem, bastante refinada e tão (ou mais) homogênea quanto a desenvolvida por Greimas, deixa o poder de análise da semiótica da Escola de Paris ainda mais atraente para os trabalhos de processamento de linguagem natural (PLN).

O termo “inteligência semiótica” que propomos, entretanto, é uma analogia realizada a partir do processo número 18/22511-0 da Fapesp que instituiu a pesquisa “Clarice.ai: uma inteligência linguística de auxílio à escrita de não ficção baseada na *web*”. Segundo Felipe Iszlaji de Albuquerque, o pesquisador responsável, o projeto pode ser definido como a utilização de epistemologia e metodologia linguística para o gerenciamento de uma atividade computacional.

O projeto de pesquisa aqui apresentado propõe a construção de uma inteligência linguística capaz de assistir em tempo real o usuário que esteja escrevendo um texto de não ficção. Ela pode ser comparada com os assistentes de correção ortográficos e gramaticais já existentes em editores de texto como o Microsoft Word e outros. No entanto, o diferencial do projeto é construir uma inteligência linguística não nos níveis ortográficos e gramaticais, mas sim no nível do estilo empregado em diferentes gêneros de não ficção.¹

0.1. Motivação

Quintilhão de *bytes* é a medida utilizada pelo blog TechJury² para representar a quantidade de dados produzida pelos seres humanos todos os dias. De acordo com eles, já estamos na casa dos 2,5 quintilhões. E os números não param de subir. Até 2025, espera-se que a humanidade produza por dia 463 *exabytes* de dados (algo equivalente a 46 milhões de hds externos do tipo mais potente, vendido na Amazon por 2 mil reais, com 10 *terabytes*). Entretanto, esse número pode estar bastante

¹ ALBUQUERQUE, F.; **Clarice.ai: uma inteligência linguística de auxílio à escrita de não ficção baseada na web**; Biblioteca Virtual da Fapesp; Disponível em: <https://bv.fapesp.br/pt/auxilios/104191/clariceai-uma-inteligencia-linguistica-de-auxilio-a-escrita-de-nao-ficcao-baseada-na-web/>; Consulta: 13/10/2020

² <https://techjury.net/blog/how-much-data-is-created-every-day/>

subestimado, uma vez que os efeitos causados pela disponibilização de internet móvel 5G em todo o planeta ainda são incalculáveis.

Como medida para as disrupções causadas pela introdução de novas tecnologias de produção e processamento de dados, basta pensarmos que antes do 4G empresas como Waze, Uber e iFood encontrariam inúmeras adversidades para triangular servidores, fornecer estimativas de geolocalização em tempo real e oferecer interface amigável e de utilização rápida e intuitiva por parte de seus usuários (que realizam pagamentos automatizados, além de terem disponível integração com outras fontes de dados, como redes sociais digitais etc.)

Entretanto, a “gula” das grandes empresas não está mais passando incólume. Em dezembro de 2020, com a demissão de Timnit Gebru³, responsável pela ética em inteligência artificial no Google, soou-se um alerta: o processamento desordenado e não gerenciado de todos esses dados já está afetando o planeta. Grandes modelos de linguagem, como o *GPT3*, endeusado pela capacidade de gerar texto automaticamente com qualidade muito alta, escondem uma dura realidade: altas taxas de emissão de gás carbônico⁴ para o treinamento das inteligências artificiais e processamento de dados de baixa qualidade, pois não curados (*fake news*, discursos de ódio, entre outros, entram no balaio do processamento descontrolado que constantemente estão enviesando os trabalhos das máquinas - conforme inúmeras denúncias da própria Gebru).

De tudo isso, a semiótica discursiva parece despontar como forte ferramenta heurística para a otimização e a curadoria dos conhecimentos processados por algoritmos de inteligência artificial, contribuindo com “Os objetivos de desenvolvimento sustentável”⁵ propostos pela ONU, além de oferecer todo um novo campo de estudo para a nossa ciência da linguagem.

³ Confira reportagem realizada pelo MIT que debate o trabalho da pesquisadora e apresenta uma leitura crítica do artigo que motivou a demissão da responsável por Ética e Inteligência Artificial no Google: <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>. Consulta: 18/01/2020.

⁴ Segundo a reportagem, um único treinamento de um modelo de IA com 213 milhões de parâmetros polui mais do que a soma da pegada de carbono de uma viagem de Nova Iorque para São Francisco, de um humano médio, de um americano médio e de um carro abastecido. Pra piorar, sabemos que os modelos são continuamente treinados, ou seja, neste exato momento existem milhares de modelos sendo treinados e causando grande impacto ambiental no planeta.

⁵ Disponível em: <https://brasil.un.org/pt-br/sdgs>. Consulta: 01/02/2021.

0.2. Objetivo

Pretendemos atualizar os trabalhos iniciados por Madeleine Arnold na década de 1980 e fornecer modelos de representação do conhecimento que possam otimizar as tarefas de processamento de linguagem natural, que fazem uso de inteligência artificial, com menor esforço computacional e oferecendo menor pegada de carbono do que os modelos existentes.

0.3. Justificativa

Apresentar um primeiro modelo de curadoria semiótica para a organização de dados é fundamental para que novos trabalhos sejam desenvolvidos em inúmeras outras áreas da semiótica, que não apenas a dos discursos verbais. Por exemplo, já nos anos 1980, Madeleine Arnold demonstrava a possibilidade de se utilizar a semiótica no processamento de imagens, citando trabalhos de Jean-Marie Floch, entre outros. Acreditamos, portanto, que nossa iniciativa serve de incentivo a mais para outros pesquisadores se debruçarem sobre os desafios contemporâneos da inteligência artificial.

Nós, linguistas e semioticistas, não precisamos ter receio e muito menos vergonha de nos aventurarmos pelos mares mais práticos de nossa disciplina. Muito pelo contrário. A linguística e a semiótica têm o ferramental necessário para assumir a dianteira na solução de inúmeros problemas computacionais. Agora, sobre “fazer teoria” versus “resolver questões práticas”, Hjelmslev, citando Bréal, em 1953, já dava grande importância ao mesmo viés de pesquisa que o nosso.

Bréal criticava na linguística o fato de ela repudiar todo e qualquer propósito prático, da mesma maneira, diz ele, que o astrônomo calcula os movimentos dos corpos celestes sem, porém, interessar-se pelas conclusões práticas que se podem tirar desses movimentos em relação às marés. Acha Bréal que o fato de visar tanto a objetivos práticos quanto teóricos nenhum dano causaria à linguística. Segundo ele, é precisamente mediante o estudo do conteúdo semântico das formas linguísticas que um tipo de linguística aplicada poderia ser criada, a par do tipo teórico.

(...) A semântica é ainda hoje objeto de muitas incertezas e discussões; todavia, os linguistas de hoje não concordam num ponto, qual seja, a importância do esforço no sentido de construir uma ciência semântica, uma doutrina do conteúdo semântico das formas linguísticas cientificamente fundamentada e realizada, o que, está visto, não apenas apresenta um interesse teórico, como ainda pode ser de importância direta para o futuro da humanidade. (1991: 104)

0.4. Contribuição

Ao final desta pesquisa, espera-se fornecer evidências do potencial heurístico da disciplina semiótica e material suficiente para demais pesquisadores classificarem o próprio corpus e desenvolverem novos modelos de inteligência artificial.

0.5. Material e método

Nosso trabalho tem como corpus mais de 2500 textos de avaliação de restaurantes publicados na plataforma digital TripAdvisor. No presente momento da pesquisa, descartamos todos os metatextos (título, data, nome do usuário etc) para nos debruçarmos apenas na avaliação textual propriamente dita (confira 2.1. Objeto “avaliação de restaurante no TripAdvisor”). Para definir e delinear nosso objeto utilizamos a epistemologia e a metodologia da semiótica discursiva, inclusive para o distinguir de outros tipos de sanção. Depois, por meio da hipótese tensiva de Claude Zilberberg, desenvolvemos uma modelagem para classificar as avaliações e a aplicamos por meio de anotação manual ao nosso corpus. Ao final, como prova de conceito, utilizamos algoritmo de *deep learning*⁶ para a realização de classificação automática.

0.6. Organização do trabalho

No primeiro capítulo, apresentamos um pouco da história da interface entre semiótica e computação. Para isso, partimos dos trabalhos de Algirdas J. Greimas e

⁶ Optou-se por não traduzir *deep learning* para “aprendizado profundo” uma vez que o termo vem se popularizando no Brasil com a grafia estrangeira - diferentemente de *machine learning*, em que o termo aprendizado de máquina já se encontra bastante disseminado.

de Pierre Maranda realizados nas décadas de 1960 e 1970 e que versam sobre a utilização da disciplina no processamento de linguagem natural. No fim desta parte, aproveitamos os debates realizados por Madeleine Arnold, a propósito do seminário sobre semiótica e inteligência artificial, para definir o norte de nossa pesquisa.

Depois, apresentamos o objeto “avaliação de restaurante no TripAdvisor” e produzimos nossas primeiras análises semióticas para revelar o modelo pressuposto pela simples existência de um enunciado daquele tipo. Isso será feito por meio do esquema narrativo canônico aplicado a um restaurante e pelo debate da classe da sanção. Neste segundo capítulo, utilizamos alguns avanços recentes da disciplina, como o modelo catenário de Bevidas, para atualizar o famoso quadrado semiótico.

Por fim, no terceiro capítulo, apresentamos brevemente o conceito por trás das análises de sentimento e apresentamos a quantificação subjetiva, conforme proposta de Luiz Tatit. Esta parte se encerra com uma breve prova de conceito em que se apresentam alguns desenvolvimentos recentes de redes neurais para exemplo de metodologia de *deep learning*.

Optamos por verter todas as citações para o português para melhor expor nossa leitura e interpretação dos originais. Por conta disso, e por não prejudicar em nada a pesquisa, dado o escopo do trabalho, utilizamos também traduções em língua portuguesa que tenham sido realizadas por terceiros.

1. HISTORIOGRAFIA DA INTERFACE ENTRE SEMIÓTICA E COMPUTAÇÃO

Neste capítulo faremos um curto e breve apanhado de algumas pesquisas realizadas entre os anos de 1960 e de 1980 para demonstrar a relação entre semiótica discursiva e ciência da computação. Nossa intenção é apresentar a proposta de Madeleine Arnold de utilizar a disciplina iniciada por Algirdas J. Greimas como forma de representação/ organização do conhecimento a ser utilizado junto às inteligências artificiais.

Para chegarmos a tal ponto, entretanto, partiremos de um texto do próprio Greimas em que, já em 1970, ele endossava a utilização de técnicas computacionais para a análise e o processamento de linguagem natural. O trecho da obra *Semiótica e Ciências Sociais* é bastante curto. Ainda assim, ele define a segmentação do texto. A nosso ver, esse é um dos primeiros desafios que o semioticista enfrenta quando se debruça sobre a utilização das ciências da computação para realização de suas análises. Afinal, devido ao percurso gerativo do sentido, muitas vezes é bastante problemático enxergar a interação entre a máquina e os níveis semionarrativos.

Após apresentarmos esse pequeno debate proposto por Greimas, no qual o fundador da Escola de Paris se vale dos trabalhos de Pierre Maranda para demonstrar a utilização da máquina, introduziremos uma breve amostra das pesquisas realizadas por esse pesquisador. Já na década de 1960, Maranda fazia utilização de técnicas computacionais para analisar os discursos produzidos por tribos brasileiras. Possivelmente, ele foi o primeiro pesquisador a aliar estruturas semionarrativas para a análise quantitativa de dados. Infelizmente, devido ao escopo deste trabalho, não poderemos nos deter muito nas pesquisas de Maranda que, até os anos 2000, ainda avançou bastante neste sentido.⁷

⁷ É interessante notar que, em 1988, Maranda evoluiu ainda mais as intersecções entre semiótica e computação por meio do artigo "Semiotics and Computers: The Advent of Semiotronics" (in **The Semiotic Web** 1987. Berlin, Boston: De Gruyter Mouton. doi: <https://doi.org/10.1515/9783110868388.507>). Devido ao escopo deste trabalho, não haverá espaço para debater os importantes pontos levantados pelo autor.

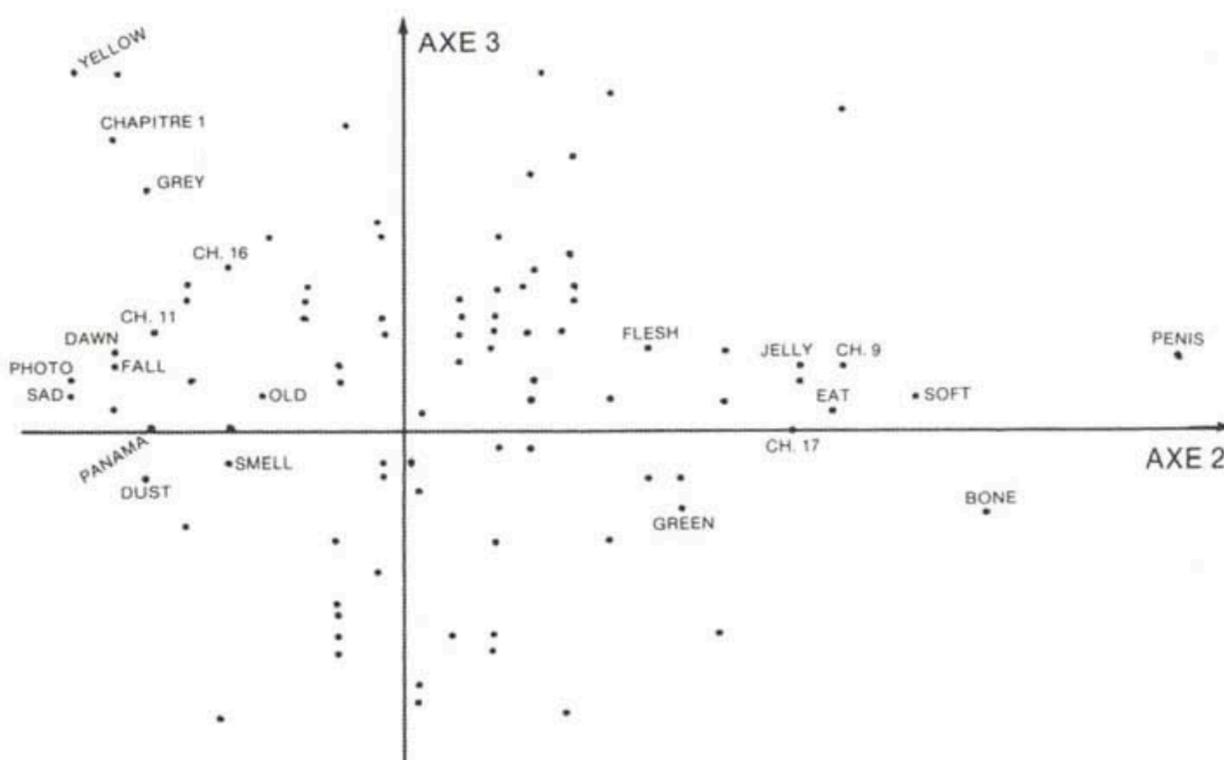
Por fim, na terceira parte apresentaremos a proposta de Madeleine Arnold para utilização da semiótica como organizadora do conhecimento a ser utilizado junto a inteligências artificiais. Antes desse trabalho, vale mencionar, a semioticista, que fora orientada em seu doutoramento pelo próprio Greimas, defendeu em 1978 a tese “La machine molle de W. S. Burroughs: étude sémiotique assistée par ordinateur des structures narratives du roman” (ou, em tradução livre “A máquina mole de W. S. Burroughs: estudo semiótico, assistido por computador, das estruturas narrativas do romance”). Um pequeno resumo deste trabalho foi publicado pela pesquisadora em razão do Segundo Congresso Internacional da Associação para Estudos Semióticos de Viena, em 1979. No artigo, Madeleine Arnold apresenta algumas evoluções em relação ao trabalho de Maranda e até mesmo dos de Greimas: ela propõe o conceito de programa discursivo, a ser depreendido a partir dos capítulos (entre outros segmentos) da obra analisada. A partir do conceito, Arnold fornece correlações quantitativas entre certas figuras discursivas que desenvolvem determinado tipo de isotopia entre os capítulos.

Em primeiro lugar, especificarei que a hipótese subjacente à escolha da análise fatorial como instrumento de análise semântica e à seleção das 94 palavras entre as mais frequentes do livro é a de que a existência de um programa discursivo é manifestado pelo surgimento de constelações lexemáticas e que a recorrência mais ou menos forte de uma constelação indica o grau de importância do programa discursivo correspondente. A análise fatorial corresponde a palavras com localizações e densidades de localização semelhantes; é, portanto, um meio de localizar constelações de palavras e os segmentos textuais que as circundam. As 94 palavras são representadas por pontos que formam uma nuvem no espaço dos 17 capítulos; as distâncias entre os pontos são uma função das frequências relativas das palavras nos capítulos. Para comodidade do estudo, a nuvem de pontos é visualizada por projeção em diferentes planos, chamados de planos fatoriais e determinados por vetores ortogonais entre si, todos passando pelo centro de gravidade da nuvem. A orientação dos vetores, denominados eixos fatoriais, é

calculada de forma a dar a imagem mais representativa possível das distâncias no espaço de origem. ⁸ (Arnold; 1984: 713-714)

Figura 1 - “Plano fatorial determinado pelos eixos 2 e 3

As palavras anotadas no plano são aquelas sublinhadas no segundo eixo pela análise fatorial. As outras palavras são simplesmente indicadas por um ponto. Uma dúzia de pontos que estavam fora do escopo da tabela não foram anotados.”⁹



Fonte: reprodução (Arnold; 1984: 719)

⁸ Tradução nossa para: “Je préciserai d'abord que l'hypothèse, sous-jacente au choix de l'analyse factorielle comme instrument d'analyse sémantique et à la sélection des 94 mots parmi les plus fréquents du livre, est que l'existence d'un programme discursif se manifeste par l'émergence de constellations lexématiques et que la récurrence plus ou moins forte d'une constellation signale le degré d'importance du programme discursif correspondant. L'analyse factorielle rapproche les mots qui ont des localisations et des densités de localisation voisines ; elle est donc un moyen de repérer des constellations de mots et les segments textuels qui englobent celles-ci. Les 94 mots sont représentés par des points qui forment un nuage dans l'espace des 17 chapitres; les distances entre les points sont fonction des fréquences relatives des mots dans les chapitres. Pour les commodités de l'étude, le nuage de points est visualisé par projection sur différents plans, dénommés plans factoriels et déterminés par des vecteurs orthogonaux entre eux et passant tous par le centre de gravité du nuage. L'orientation des vecteurs, appelés axes factoriels, est calculée de telle façon qu'elle donne l'image la plus représentative possible des distances dans l'espace origine.”

⁹ Tradução nossa para: “Plan factoriel déterminé par les axes 2 et 3. Les mots notés dans le plan sont ceux qui sont soulignés sur le deuxième axe par l'analyse factorielle. Les autres mots sont simplement indiqués par un point. Une douzaine de points qui sortaient du cadre du tableau n'ont pas été notés.”

Todo o trabalho da pesquisadora centra-se na obra *Soft Machine*, de William Burroughs, publicada em 1961. Segundo Arnold, o título faz referência a uma máquina mole que representa tanto o corpo humano quanto a humanidade. Para a semiótica, o livro foca a sexualidade e a violência, abordadas tanto individual quanto coletivamente, além dos mecanismos de controle que são realizados por meio de manipulação ideológica e de inúmeros tipos de repressão.

Além de uma forte análise semi-narrativa, que localiza actantes e a sintagmatização de um quadrado semiótico orientado entre 'vida' e 'morte', Arnold pincela no artigo o emprego da técnica estatística conhecida como análise fatorial para construir, a partir da distribuição de palavras-chaves em todos os capítulos do livro, as principais constelações lexemáticas (acepção retirada de Hjelmslev para se referir a algo que hoje conhecemos como nuvem de palavras). Na figura 1 temos uma pequena amostra da representação gráfica.

Junto a essa pequena introdução, espera-se que a leitura das próximas seções do capítulo evidenciem a consistência das pesquisas que aliaram semiótica e ciências da computação. Mais do que isso, esperamos que também sirva de alerta: é preciso olhar para o passado da disciplina, pois muito trabalho de grande envergadura já foi feito e pode ser utilizado para resolver inúmeros problemas computacionais da contemporaneidade - principalmente aqueles que dizem respeito ao processamento de linguagem natural.

1.1. Algirdas Greimas e Ciências da Computação

Criada em fins dos anos 1960 pelo linguista lituano Algirdas Julien Greimas, a semiótica da Escola de Paris sempre se manteve atenta aos desenvolvimentos tecnológicos (fato que buscaremos evidenciar em todo este primeiro capítulo). Já em abril de 1970, durante o Simpósio Internacional de Literatura Técnica, em Palermo, na Itália, o antropólogo Pierre Maranda apresentou os primeiros resultados de sua pesquisa, que aliava o modelo narrativo canônico e técnicas quantitativas de análise de corpora.

Antes de apresentarmos o trabalho realizado por Maranda junto a um conjunto de relatos mitológicos de tribos do grupo linguístico jê (localizados na região do Mato Grosso), e que será feito na seção seguinte a esta, cabe apresentar um breve apanhado sobre o pensamento de Greimas em relação à Ciência da Computação.

Um primeiro traço comum, fundamental, apareceu quando se perguntou sobre a possibilidade de substituir qualquer abordagem intuitiva por um conjunto de critérios objetivos, no momento da segmentação do texto. A tendência extrema, que visa a tornar possível a leitura automática do texto foi representada por Pierre Maranda que propõem refletir em termos de autômato, simular inteiramente o comportamento cognitivo do pesquisador para tentar em seguida reproduzi-lo através da máquina. ([1976] 1981: 171)

Com as palavras acima, Greimas apresenta o problema que se colocava naquela época e que, durante toda esta pesquisa, foi colocado para nós: como tornar a prática do pesquisador suficientemente objetiva ao ponto de ela poder ser reproduzida por um computador? Como programar o comportamento analítico do semiótico para ele poder ser totalmente reproduzido por uma máquina? Não por acaso, o nome da seção em que o excerto acima se encontra é "A segmentação do texto".

Um dos principais problemas de intersecção entre semiótica e ciência da computação diz respeito às operações computacionais se realizarem junto ao nível de superfície de um texto, junto às letras. Além disso, um computador não programado é 100% burro. Ele só sabe o que o programador ensina a ele e, sendo assim, é preciso dizer exatamente quais palavras, quais combinações de palavras, etc., representam aquilo que o analista sabe intuitivamente: que uma manipulação está sendo realizada, que uma performance não foi cumprida e por aí vai.

Mais do que isso, para o computador, a elasticidade da língua é um grande problema. Pois, apesar de, como diz Tatit, o "bom funcionamento da nossa língua de todos os dias depende de uma alternância até certo ponto equilibrada entre a redução de conteúdos em palavras-chaves e sua expansão em argumentos discursivos" (2019: 163); essa característica introduz ambiguidades que o computador sozinho não é capaz de solucionar. Como ensinar a um computador que

os pronomes 'esse', 'este' e 'aquele' se referem a termos do próprio texto e que sempre variam de acordo com ele?¹⁰ Greimas avança mais:

As divergências apareceram, creio eu, no momento em que se colocou a questão de saber o que se segmenta na realidade. Existem, de fato, duas segmentações possíveis: por um lado, a segmentação do texto como ele aparece em língua natural, a segmentação do discurso em suas partes constitutivas, e, por outro, a segmentação, também possível, do texto considerado como uma narrativa, em suas unidades narrativas. Fiquei com a impressão de que esta distinção permaneceu bastante flutuante, e que um perigo de confusão ameaçava aparecer, pelo menos no momento da generalização eventual de nosso saber-fazer. ([1976] 1981: 171)

Ou seja, mais do que a elasticidade discursiva, os níveis textuais também são um problema para o analista. Afinal, as unidades narrativas são bastante complexas e não dizem respeito, estritamente, apenas ao enunciado que se avalia¹¹. O esquema narrativo canônico de Greimas constitui-se de manipulação, competencialização, performance e sanção. Entretanto, muitas dessas partículas narrativas podem estar apenas pressupostas em um texto. Tal fato parece gerar muitos problemas para o semiótico quando ele resolve programar um computador.

Segundo MATTE e LARA, o nível narrativo é composto de funtivos e suas funções, "relações lógicas, extemporais e praticamente esvaziadas de conteúdo figurativo e temático" (2008; não paginado¹²). Para a pesquisadora, um bom exemplo seria a história do patinho feio.

Era uma vez um cisne maravilhoso e muito feliz, que escondia por trás de sua exuberância uma história bastante dolorosa. Quando chegou ao lago onde agora vive com seus companheiros, ele sequer poderia imaginar que seu percurso de dor estava prestes a acabar. (MATTE e LARA; 2008: np)

A história está claramente sendo contada do fim para o começo. O patinho já se tornou cisne e toda dor parece ter sido superada. Entretanto, o "que garante que este é o final da história, e não o começo, mesmo que sejam nossas primeiras

¹⁰ Um dos motivos de nossa aposta nas técnicas de *deep learning* (no lugar de aprendizados de máquina mais convencionais) advém, entre outros, de desenvolvimentos recentes que parecem já ter solucionado tal problema.

¹¹ O que ficará mais claro nos próximos capítulos (quando apresentarmos uma proposta de análise da enunciação pressuposta em uma avaliação de restaurante na plataforma TripAdvisor).

¹² A partir de agora np.

palavras ao iniciar esse relato, é o nível narrativo." (MATTE e LARA; 2008: np) Para resumir: a ordem narrativa é sempre pressuposta em um texto. Por exemplo, para fazer (realizar uma performance, uma transformação em um estado próprio ou de outrem), um sujeito antes precisa querer (ser manipulado para perceber que está separado de um objeto de valor do qual ele necessita), depois precisa poder e saber (ter competência para realizar o ato) e por fim realizar a performance que poderá ser sancionada (negativa ou positivamente) de acordo com a existência de um contrato. Bastante informalmente, tentamos apresentar algumas dificuldades enfrentadas pelo semiótico que resolve programar uma máquina para reproduzir uma prática que ele realiza com bastante tranquilidade. Entretanto, Greimas não se contenta apenas com as problematizações já apresentadas e resolve aprofundar o tema ainda mais.

Parecia-me assim que o projeto de Pierre Maranda - que subscrevo inteiramente - não deixava aparecer suficientemente, no nível de suas aplicações, a distinção entre o que se chama a linguagem lógico-semântica e a linguagem-máquina. Sabe-se, por exemplo, que mesmo os textos matemáticos, aparentemente mais formalizados, para se submeterem ao tratamento do computador, devem sofrer uma reformulação mediadora chamada linguagem-máquina. As coisas, nesse nível, não são muito claras: por um lado, a linguagem-máquina exprime em termos unívocos o que fica ainda implícito no texto e parece, assim, mais poderosa do que a lógica "natural" de um matemático. (GREIMAS; [1976] 1981: 171)

Já em 1970, Greimas reconhece o poder da linguagem de máquina, aquela composta apenas por 0s e 1s (abaixo apresentaremos uma breve história da evolução dessas linguagens) e que apenas os raros programadores que trabalham com linguagem de baixo nível conhecem bem. Como observação, cabe mencionar que em um artigo intitulado "As 7 linguagens de programação que você deve aprender em 2020"¹³, nenhuma delas é linguagem de máquina, todas são linguagem de alto nível (os termos alto e baixo nível, em programação, não carregam juízo de valor, apenas reportam a relação entre linguagens mais fáceis de compreender por um humano, pois mais próximas da sintaxe e da semântica, e a linguagem binária). Entre uma e outra, existem tradutores (compiladores, interpretadores) que realizam a conversão e permitem que um código seja realizado pela máquina.

¹³ Disponível em: <https://computerworld.com.br/carreira/as-7-linguagens-de-programacao-que-voce-deve-aprender-em-2020/>. Consulta: 01/02/2021

Mas, por outro lado, esta linguagem-máquina obedece apenas às exigências pragmáticas da utilização dos computadores: deste ponto de vista, ela é umas das linguagens documentárias possíveis e, conseqüentemente, inferior à linguagem lógica. A partir daí, parece que se impõe a necessidade de examinar e de precisar o estatuto da linguagem documentária a ser utilizada com vistas à segmentação e à leitura automática dos textos. (GREIMAS; [1976] 1981: 171)

Se houve um primeiro assombro, por parte de Greimas, ao perceber que a linguagem de máquina seria capaz de dar conta de representar muito mais fenômenos do que a "'lógica' natural de um matemático", o assombro rápido se esvanece. Logo, surge a percepção de que tal linguagem é uma "simples" notação a serviço da prática do programador - e que muitas outras existem. Tomando apenas as palavras de Greimas, somos levados a crer que o semiótico ainda não tinha conhecimento das linguagens compiladas e/ou interpretadas (que são as mais utilizadas atualmente, como mencionamos acima).

Cada computador, seja antigo ou moderno, tem associado a si um conjunto de instruções que seu hardware é capaz de realizar. Esse conjunto constitui a linguagem de máquina desse computador e cada uma de suas instruções é expressa por seqüências de bits. Os programadores de computadores primitivos só contavam com essa linguagem para escrever seus programas, fossem esses estabelecidos por conexões eletrônicas ou escritos em cartões perfurados. (MOKARZEL e SOMA; 2008: 35)

Da linguagem de máquina escrita em binário, avançou-se para o sistema decimal. Ainda assim, além do trabalho maçante, com falta de clareza e muitos erros, em que um bit 1 comumente podia ser trocado por 0, permanecia a obscuridade entre as relações mantidas pelos números e pela ação que eles deveriam representar.

No início dos anos 50, aparecem as chamadas linguagens de montagem ou linguagens Assembly, ou simplesmente Assembly's, com a finalidade de expressar as instruções de máquina de forma mais clara. Os códigos de operação dessas instruções passaram a ser expressos por mnemônicos, e as informações sobre os operadores por nomes que lembrassem seu significado. Endereços de memória poderiam ser referenciados por tais nomes e isso recebeu o nome de endereçamento simbólico. (MOKARZEL e SOMA; 2008: 39)

Atualmente, linguagens com estruturas mais bem definidas, que se aproximam muito da sintaxe e da semântica das línguas naturais (ou seja, com fácil interpretação por seres humanos) foram desenvolvidas. O poder de tais linguagens computacionais aumentou muito a capacidade de produção de programas, bem como na consolidação de novos e diferentes paradigmas para se resolver uma tarefa. São conhecidas como linguagem de alto nível, ou apenas linguagem de programação: Cobol, Pascal, C, C++, Prolog, Java, Python, entre outras.

Para que programas escritos nessas linguagens de programação possam ser executados num computador, é necessário um processo de tradução bem mais complexo para linguagem de máquina do que aquele envolvendo programas escritos em Assembly, pois a estrutura dessas linguagens é totalmente diferente daquela das linguagens de máquina. (MOKARZEL e SOMA; 2008: 43)

Nossa intenção nesta seção, muito longe de esgotar o tema, foi apenas de apresentar os conhecimentos e as inquietações de Greimas com as Ciências da Computação já em 1970. Muitas pesquisas, inclusive de cunho historiográfico, ainda são necessárias para que se compreenda todos os imbricamentos possíveis entre semiótica e linguagem de programação. Em relação aos desenvolvimentos das questões relativas à segmentação do texto, todas as seções e capítulos desta dissertação continuarão a se debruçar e aprofundar o tema, pois ele diz respeito fundamentalmente à questão da representação do conhecimento pela semiótica.

1.2. Pierre Maranda e a Análise Computacional do Mito

Greimas, na seção anterior, queixava-se a respeito de as pessoas estarem, aparentemente, divergindo sobre a segmentação do texto. Então, ele explica que existe uma segmentação discursiva e outra semionarrativa. Entretanto, o linguista não desaprova a utilização apenas de temas e figuras como forma de produzir análises bastante pertinentes. Principalmente quando manejadas por um antropólogo conhecedor das análises quantitativas e da programação. Antes de chegarmos em Pierre Maranda, propriamente dito, cabe explicar melhor a segmentação do texto apenas pelo nível discursivo. Atividade que foi realizada

inúmeras vezes pelo antropólogo, inclusive com a produção de dados etnográficos a partir da observação de populações urbanas no Canadá¹⁴.

Ótimo exemplo sobre tal tipo de segmentação são os trabalhos realizados pela brasileira Diana Luz Pessoa de Barros junto aos discursos das redações de vestibular. Apresentada como "A festa do discurso. Teoria do discurso e análise de redações de vestibulandos", a tese de livre-docência da professora uspiana, já em 1985, apresentava a possibilidade de realizar uma inferência contextual de um texto a partir das classes sociais de que ele se orienta.

Temas e figuras constituem a semântica discursiva: os temas são os conteúdos semânticos tratados de forma abstrata, e as figuras, o investimento semântico-sensorial dos temas. Os temas e figuras são determinados sócio-historicamente e trazem para os discursos o modo de ver e de pensar o mundo de classes, grupos e camadas sociais, assegurando assim o caráter ideológico desses discursos. (BARROS; 2009: 352)

Munida das redações realizadas pelos alunos ingressantes na Universidade de São Paulo, e que tinha como assunto "festa de desconhecidos", bem como das respectivas informações socioeconômicas dos candidatos, Pessoa de Barros traçou os principais temas desenvolvidos pelas classes médias: o da ascensão social e o da família. Segundo a pesquisadora, 80% dos vestibulandos faziam parte dessa classe.

O tema da família é, na verdade, o da passagem, em geral mal sucedida, à idade adulta, ou seja, passagem da dependência familiar infantil e adolescente à independência do adulto. Nas redações, o jovem das camadas médias sai do aconchego do lar, levado por um amigo, e para fazer novas amizades que, aos poucos, libertem-no do estreito círculo familiar. (BARROS; 2009: 353)

No caso acima, a pesquisadora relata que grande parte dos vestibulandos dizia se decepcionar, de alguma forma, com os amigos, e decidir voltar aos braços da família. Em uma outra opção de texto, também presente no corpus de Pessoa de Barros, o aluno introduzia os novos colegas na vida familiar dele - inclusive com

¹⁴ Confira o artigo "Semantografia do domínio trabalho entre as cidades alta e baixa de Quebec". (MARANDA, P.; **Sémantographie Du Domaine "travail" Dans La Haute-ville Et Dans La Basse-ville De Québec**; *Anthropologica* 20, no. 1/2 (1978): 249-92. Consulta: 18/01/2021. doi:10.2307/25605007.)

relatos de casamento. Entretanto, Pessoa de Barros também define outro grande núcleo temático a que as redações oriundas das classes médias fazem uso: “vida e festa de rico”

No tema da ascensão social, por sua vez, a festa tem belas mulheres, muito luxo, pratos finíssimos, canapés, bebidas estrangeiras, mansão no Morumbi, piscina, um batalhão de empregados – governanta, copeira, garçons, porteiro, mordomo, manobrista. Não são festas com cachorro-quente, sanduíches ou churrasco, no fundo do quintal. São festas das novelas da televisão, que o vestibulando, das camadas sociais médias, não frequenta e que figurativiza com dificuldade, pela falta de mais informações, mas de que gostaria de participar. (BARROS; 2009: 354)

O trabalho de Pessoa de Barros é munido, também, de análises sociológicas, como as de Décio Saes. A partir delas, a semioticista faz inferências como as de que existe uma certa "consciência média" que caracteriza-se, entre seus traços mais marcantes, por: "o medo da proletarização", "a ideologia da competência" e o "desejo da ascensão social". Nosso escopo não é analisar os quadros teóricos externos à semiótica de que faz uso Pessoa de Barros. Coisa que, não há dúvida, renderá ainda muita pesquisa. Distribuímos acima alguns trechos do trabalho brasileiro apenas para apresentar coisas que já eram feitas e que mantêm afinidade ao nosso trabalho. Como resumo, Pessoa de Barros demonstra a pertinência em se correlacionar dados socioeconômicos com textos produzidos por sujeitos de diferentes faixas demográficas (falaremos mais sobre os princípios semióticos postulados pela pesquisadora no próximo capítulo). Entretanto, à guisa de importância da obra, vale mencionar um trabalho apresentado recentemente pelo professor Fiorin.

O excerto, que será apresentado a seguir, precisa de um pouco de contexto: a tese de Pessoa de Barros, apresentada no concurso de livre-docência, é composta de duas partes. A primeira é bastante conhecida, pois foi publicada com o título "Teoria do Discurso. Fundamentos semióticos" (1988) e se tornou um dos principais manuais para quem se inicia na área - pois apresenta uma visão holística de todo o edifício erigido a partir de Greimas. A segunda parte, em que a pesquisadora apresenta as relações socioeconômicas e as análises semióticas, nunca foi publicada.

a não publicação da integralidade da tese fez esse livro ser visto como uma apresentação da semiótica, deixando, a meu ver, na sombra, sua principal contribuição teórica: o exame das relações entre discurso e contexto sócio-histórico, ou seja, uma proposta de integração do que era chamado análise interna e análise externa. Isso era, na época, uma proposta inovadora, pois, nos círculos semióticos, tinha-se uma ideia um tanto quanto redutora do princípio de imanência e, portanto, só a análise interna tinha lugar. Aliás, é inovadora até hoje em termos de semiótica no mundo. (FIORIN; 2020: 13)

Entretanto, o que Pessoa de Barros muito bem demonstra, já era realizado (guardadas as proporções), pelo menos desde a década de 1960, de forma computadorizada, por Pierre Maranda. Pesquisador que, conforme vimos acima, tinha total endosso do próprio Greimas. Infelizmente, não existem dados referentes à apresentação realizada pelo antropólogo no congresso de Literatura Técnica de Palermo (a não ser em caso de falha nossa). Entretanto, dos inúmeros trabalhos produzidos pelo pesquisador, na altura do congresso, um deles parece resumir bem o que Pierre Maranda estava iniciando.

Em “Formal analysis and inter-cultural studies” (Análise formal e estudos interculturais, em tradução livre), Maranda demonstra a metodologia que empregou em inúmeros de seus trabalhos. No artigo, estão disponíveis alguns resultados obtidos a partir das análises dos mitos de tribos do grupo linguístico jê (localizados na região do Mato Grosso). O antropólogo, se pudermos nos reportar apenas ao artigo de 1967, correlacionava quantitativamente a circulação de temas e figuras em tribos brasileiras.

'O cru e o cozido' analisa o relato bororo sobre a origem do fogo por meio de referências cruzadas entre oito mitos indígenas norte-americanos e 179 sul-americanos, entre os quais 38 são Jê (contra 135 que constituem o objeto desta análise). Apesar da sobreposição parcial de dados e algumas conexões metodológicas, o presente estudo difere do de Lévi-Strauss em vários aspectos. Por um lado, é exaustivo porque nenhum nível estratégico é preferido e o ponto de partida é coextensivo à totalidade dos textos; então, a pesquisa se apóia na análise quantitativa (frequência), bem como na qualitativa (contingência), enquanto Lévi-

Strauss se preocupa apenas com esse último tipo de investigação. (MARANDA; 1967: 8)¹⁵

Maranda marca sua posição (e filiação) ao estruturalismo de Lévi-Strauss ao mesmo tempo em que se distancia dele por ambicionar, aparentemente, fins mais amplos que os do pai da antropologia estrutural. De acordo com o texto, Pierre Maranda parte de um corpus composto de 135 relatos colhidos entre quatro sociedades pertencentes ao grupo linguístico jê: Apinajés, Xerente, Timbiras Orientais e Caiapós. Para realizar o trabalho, Maranda (1) editou o corpus, anotando-o sintaticamente; (2) construiu um dicionário para 'validar' os termos; (3) restringiu o corpus a um mesmo grupo, para poder realizar a análise comparativa; e (4) partiu de categorias significantes aos nativos para realizar um trabalho mais profundo.

Deve ser mencionado neste ponto que não há atalhos para a análise formal. No que diz respeito aos computadores, eles são, sem dúvida, instrumentos poderosos, mas a validade dos dados de saída depende inteiramente daqueles da entrada. Não há como alguém obter resultados confiáveis a partir de dados pobres, nem como evitar o trabalho árduo de preparar as entradas. Como ficará claro no decorrer desta apresentação, os computadores auxiliam na análise de textos, mas o tempo e a energia que economizam é apenas o retorno de um investimento preliminar do pesquisador. As decisões oriundas do baixo investimento na preparação dos dados devem ser pagas em termos de resultados piores em razão exponencial para a quantidade de tempo economizada na primeira fase. (MARANDA; 1967: 8)¹⁶

Maranda dá aula, ainda hoje, para qualquer aspirante à Ciência de Dados: computadores não fazem milagres. Portanto, após a introdução, ele apresenta a metodologia científica utilizada na preparação do corpus. Inclusive com a utilização

¹⁵ Tradução nossa para: "Le cru et le cuit analyzes the Bororo account of the origin of fire by cross-references to eight North American Indian myths and 179 South American ones among which 38 are GE (against 135 forming the object of this analysis). Notwithstanding partial overlapping of data and some methodological connections, the present study differs from Lévi-Strauss's in several respects. For one thing, it is exhaustive in that no strategic levels are singled out and that the threshold is coextensive with the full texts; then, it rests on quantitative (frequency) as well as qualitative (contingency) analysis whereas Lévi-Strauss is concerned only with the latter type of investigation."

¹⁶ Tradução nossa para: "It must be mentioned at this point that there is no short-cut to formal analysis. As far as computers are concerned, they are doubtless powerful instruments, but the validity of their outputs depends entirely on that of inputs. There is no way in which one could get reliable results from poor data, nor is there any to avoid the painstaking work to prepare inputs. As it should become clear in the course of this presentation, computers help analyze texts, but the time and energy they save is only the return of a preliminary investment by the investigator. Decisions to spare in the preparation of the data must be paid for in terms of poorer results in exponential ratio to the amount of time saved in the first phase."

de táticas ainda hoje atuais: sentenças na voz passiva foram convertidas em voz ativa, frases complexas foram divididas em sequências concatenadas por conjunções e depois normalizadas por meio da inserção de especificadores (Maranda; 1967: 9). Atividades que, a uma primeira vista, lembram muitos os trabalhos realizados por meio de lematizadores. Depois disso, por meio de um programa chamado KWIC ('Key Words in Context', ou 'palavras-chaves dentro de um contexto', em tradução livre), Pierre Maranda construiu um dicionário com 2736 entradas em que as palavras eram apresentadas a partir do contexto em que estavam inseridas dentro dos relatos míticos, seja por meio da quantidade de ocorrências ou em ordem alfabética. Algo aparentemente muito semelhante ao que os concordanciadores realizam hoje em dia: fornecimento de uma quantidade de palavras antes e outra depois do termo selecionado para ser definido.

No contexto da análise automática, um dicionário é essencialmente um conjunto de categorias às quais o computador reduz a multiplicidade de dados para fins descritivos e sintéticos. Essa normalização de segunda ordem é então usada em fases posteriores da pesquisa para análises quantitativas e qualitativas. Essas categorias podem ser construídas a priori ou podem surgir dos próprios documentos. O programa KWIC é mais útil para a abordagem 'êmica', pois imprime uma concordância detalhada. (MARANDA; 1967: 11)¹⁷

A partir dessa atividade, o pesquisador apresenta os principais temas encontrados e representados no dicionário (figura 2, abaixo). No artigo, Pierre Maranda menciona a existência de uma gramática subjacente à lista de temas, a qual seria construída por meio de interpretações sociológicas da antropologia estrutural. "O dicionário está principalmente preocupado com representações coletivas" (MARANDA; 1967: 14)¹⁸. Tal fato, de imensa importância para a análise automatizada de corpus, já era denunciado por Greimas em *Semiótica e Ciências Sociais*.

A presença - ainda parcial - do contexto situacional implícito deve conduzir-nos à constituição de um dicionário mitológico, isto é, ao estabelecimento, para cada

¹⁷ Tradução nossa para: "In the context of automatic analysis, a dictionary is essentially a set of categories to which the computer reduces the multiplicity of data for descriptive and synthetic purposes. This second-order normalization is then used in ulterior phases of research for quantitative and qualitative analyses. Such categories may be constructed a priori, or they may emerge from the documents themselves. The KWIC program is most helpful to the 'emic' approach since it prints out a detailed concordance."

¹⁸ Tradução nossa para: "the dictionary is mainly concerned with collective representations."

cultura determinada, de um inventário lexemático feito de palavras ou de expressões que, no momento da análise concreta, parecem "significativos" pelo fato de que, por exemplo, manifestavam actantes ou funções dentro do esquema narrativo antevisto. Tal dicionário, evidentemente, só deveria ser estabelecido para ser transformado, depois da análise, em código semântico de um determinado universo cultural. Teoricamente, só a posse de tal código semântico permitiria a análise semântica de uma narrativa particular. Na sua ausência, os resultados de nossas análises permanecem sempre hipotéticos e levantam problemas de verificação muito complexos. A distância entre as exigências teóricas e nossas possibilidades práticas é, como vemos, considerável. (GREIMAS; [1976] 1981: 174)

Figura 2 - Temas apreendidos a partir da construção de um dicionário mitológico

01 COSMOLOGICAL, SKY	44 SOCIOLOGICAL, AFFINE
02 COSMOLOGICAL, EARTH	45 SOCIOLOGICAL, CEREMONIAL RELATIVE
03 COSMOLOGICAL, FIRE	46 SOCIOLOGICAL, UNSPECIFIED RELATIVE
04 COSMOLOGICAL, WATER	47 SOCIOLOGICAL, OWNERSHIP
05 MOTION, TRANSITIVE [a subject moves an object]	48 SOCIOLOGICAL, CRIME
06 MOTION, INTRANSITIVE [a subject moves itself]	49 SOCIAL BEHAVIOR, SOCIABILITY [friendly relations]
07 MOTION, CYCLICAL [return, etc.]	50 SOCIAL BEHAVIOR, COMMUNICATION
08 MOTION, PRESENCE [appearance, contact with, etc.]	51 SOCIAL BEHAVIOR, COMPETITION
09 MOTION, ABSENCE [departure, disappearance, etc.]	52 SOCIAL BEHAVIOR, HOSTILITY
10 MOTION, STOP	53 SOCIAL BEHAVIOR, AGGRESSION
11 SPACE, PLACE	54 SOCIAL BEHAVIOR, DESTRUCTION
12 SPACE, POSITION	55 SOCIAL BEHAVIOR, DOMINANCE
13 SPACE, SETTLEMENT	56 SOCIAL BEHAVIOR, SUBMISSION
14 TIME, GENERAL [unspecified references to time]	57 SOCIETY, TRIBE [larger social units : other tribes, <i>civilizados</i> , etc.]
15 TIME, DAY	58 SOCIETY, GROUP [villagers, bands, tribesmen, etc.]
16 TIME, NIGHT	59 SOCIETY, CLASS [age classes, clans, men's societies, etc.]
17 QUANTITATIVE, GENERAL	60 SOCIETY, CHIEF
18 QUANTITATIVE, SIZE AND SHAPE	61 SOCIETY, MALE [to be completed by cross-references to tag 63]
19 PLANT, WHOLE	62 SOCIETY, FEMALE [entries in 43, 44, 45, 46, 58, 59, and 60]
20 PLANT, PART	63 SOCIETY, LONELINESS [individual isolated from his people]
21 ANIMAL, WILD	64 SOCIETY, UNDIFFERENTIATED [suckling, infant, etc.]
22 ANIMAL, DOMESTICATED	65 RITUAL, CEREMONY
23 ANIMAL, AMPHIBIOUS	66 RITUAL, PARAPHERNALIA
24 ANIMAL, FISH	67 RITUAL, SIGN [dream, vision, etc.]
25 ANIMAL, BIRD	68 SUPERNATURAL, GOD [reference to Christian conceptions of God]
26 ANIMAL, INSECT	69 SUPERNATURAL, CELESTIAL
27 BODY PART, GENERAL [skin, bones, blood, etc.]	70 SUPERNATURAL, MAGIC
28 BODY PART, HEAD	71 SUPERNATURAL, CULTURE HERO
29 BODY PART, LIMB	72 PSYCHOLOGICAL, PASSIVE [state of mind, passive feeling]
30 BODY PART, ORGAN	73 PSYCHOLOGICAL, ACTIVE [determination, dynamic feelings]
31 BODY PART, SOUL	74 COGNITION, PERCEPTUAL, SIGHT
32 BODY PART, TRUNK	75 COGNITION, PERCEPTUAL, HEARING
33 PHYSIOLOGICAL, DISEASE	76 COGNITION, PERCEPTUAL, SMELL AND TOUCH
34 PHYSIOLOGICAL, HEALTH	77 COGNITION, PERCEPTUAL, TASTE
35 PHYSIOLOGICAL, SECRETIONS	78 COGNITION, PERCEPTUAL, KINESTHETIC [heavy, fast, strong, etc.]
36 PHYSIOLOGICAL, ALIMENTATION	79 COGNITION, PERCEPTION, KNOWLEDGE [more abstract cognitive processes]
37 PHYSIOLOGICAL, DEATH	80 COGNITION, DELIBERATION AND DECISION
38 PHYSIOLOGICAL, REST	81 FACTS [all that is reported as "relation", "narration", etc.]
39 PHYSIOLOGICAL, BIRTH	82 TASK
40 PHYSIOLOGICAL, SEX	83 TECHNOLOGY, NATIVE, GENERAL
41 PHYSIOLOGICAL, LIFE	84 TECHNOLOGY, ORNAMENT
42 PHYSIOLOGICAL, NEED	85 TECHNOLOGY, TRAVEL
43 SOCIOLOGICAL, KIN	86 TECHNOLOGY, COOKING [cf. Lévi-Strauss 1964]
	87 TECHNOLOGY, WHITE MAN
	88 OPERATION INCLUSION [also, and, either]
	89 OPERATION EXCLUSION [negation in general]
	90 OPERATION IMPLICATION [accordingly, if, so that, on condition, etc.]
	91 OPERATION EVALUATION [good, bad, more, less, too much, etc.]
	92 OPERATION EVENTUAL [might, perhaps, possible, would have, etc.]
	93 OPERATION CAUSALITY [by chance, in order to, etc.]
	94 OPERATION ANTERIORITY [before, already, previous, etc.]
	95 OPERATION SIMULTANEITY [while, at the same time, meantime, etc.]
	96 OPERATION POSTERIORITY [after, behind, then, etc.]
	97 OPERATION COMPARISON [alike, same, resemble, etc.]
	98 TRANSFORMATION, INTRINSIC [actor changes itself into]
	99 TRANSFORMATION, EXTRINSIC [actor changes an object into]

Fonte: reprodução (MARANDA; 1967: 12-13)

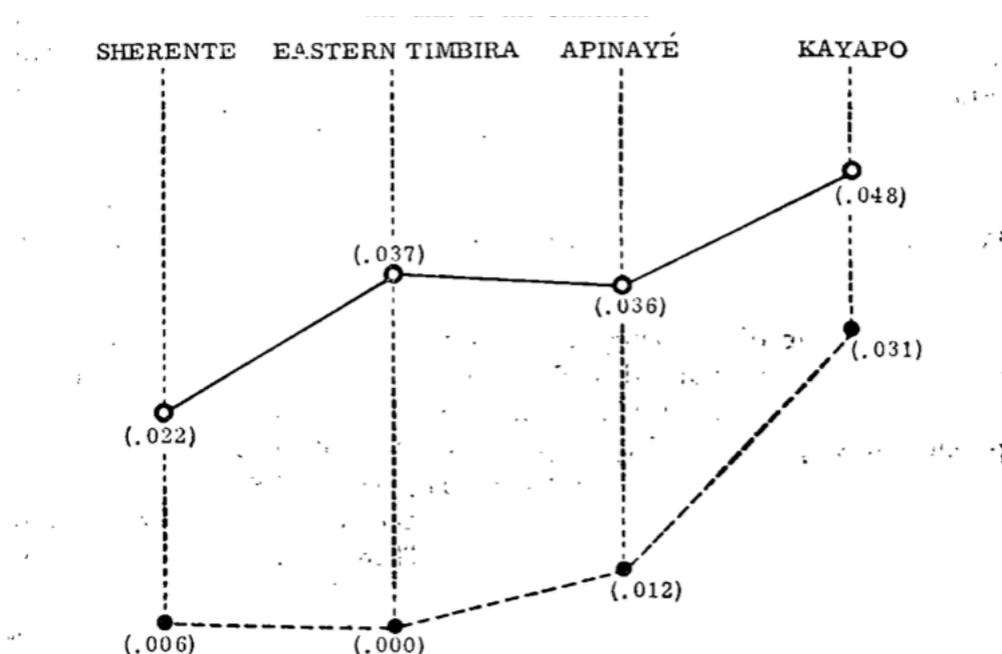
Infelizmente, devido ao escopo do presente trabalho, não podemos nos prender a minúcias da pesquisa de Pierre Maranda, sem dúvida ainda hoje muito atual: não são poucos os aprendizados que os cientistas de dados, que trabalham com

processamento de linguagem natural, poderão obter através da recuperação dos trabalhos do antropólogo, bem como os semioticistas que queiram enveredar pela área. Entretanto, à guisa de encerramento, é importante apresentar um pequeno exemplo de resultado oferecido por Maranda para que sintetizemos tudo o que foi exposto aqui.

Figura 3 - "Jaguar e Tapir na mitologia jê

Linha reta = Jaguar; linha tracejada = Tapir; em centésimos de 1%;

a unidade é a sentença"¹⁹ (MARANDA; 1967: 15)



Fonte: reprodução (MARANDA; 1967: 15)

Na figura 3, podemos ver a representação gráfica da presença de dois atores nos relatos míticos de diferentes povos indígenas. Segundo Maranda, podemos depreender do gráfico que a figura mítica do Jaguar (onça), representado usualmente como "o dono original do fogo domesticado", é forte entre os quatro povos pesquisados. Por outro lado, a presença do ator Tapir (anta), normalmente definido pela /força/ e por características de /amante adúltero/, só é marcante entre

¹⁹ Tradução nossa para: "Jaguar and Tapir in Gê mythology. Plain line = Jaguar; broken line = Tapir; in hundredths of 1%; the unit is the sentence."

os relatos dos Caiapós. Maranda nota que entre os apinajés, por exemplo, quando citado, o Tapir desperta interesse e é definido por sua /aparência monstruosa/ em uma frase e por sua /força/ em sete. Entretanto, a função de /amante adúltero/ cabe ao crocodilo entre aquele povo. Em todos os relatos (com forte cunho moralizador), no entanto, os atores que recebem o revestimento temático de /amante adúltero/ acabam sendo mortos pelos indígenas, como vingança pela contravenção.

1.3. Madeleine Arnold e Inteligência Artificial

Apesar do trabalho bastante inovador, sobre o qual discorreremos na introdução deste capítulo, e que merece ser recuperado (principalmente para entendermos todas as consequências por trás da proposta do conceito de programa discursivo de Madeleine Arnold), nosso interesse sobre a pesquisadora gira em torno de um relatório publicado em 1988, sobre um seminário que teve como tema Semiótica e Inteligência Artificial. O evento foi realizado no grupo de estudos conduzido por Greimas na Escola de Altos Estudos em Ciências Sociais²⁰ e contou com 10 sessões, aproximadamente, com uma média de duas horas cada, em que Jean Zeitoun, então responsável pelo Centro para Ciência da Computação e Metodologia em Arquitetura (CIMA), também na França, apresentou o estado da arte da Inteligência Artificial naqueles tempos.

Os objetivos do workshop são esclarecer as semelhanças e diferenças que existem entre a semiótica e a inteligência artificial (IA) em sua abordagem da significação e determinar as origens dessas divergências. Esta comparação das duas disciplinas é baseada em uma introdução seletiva à IA. Inclui, por um lado, a economia geral dos modos de representação e processamento do conhecimento e, por outro, duas áreas específicas da IA: compreensão da linguagem natural (NLU) e compreensão da imagem (IU). (ARNOLD; 1988: 667)²¹

²⁰ No francês: École des hautes études en sciences sociales (EHESS).

²¹ Tradução nossa para: "The aims of the workshop are to clarify the similarities and differences that exist between semiotics and artificial intelligence (AI) in their approach to signification, and to determine the origins of these divergences. This comparison of the two disciplines is based on a selective introduction to AI. It includes on the one hand the general economy of the modes of representation and knowledge processing, and on the other two specific areas of AI: natural language understanding (NLU) and image understanding (IU)."

Em nossa tradução, mantivemos os termos conforme a escrita de Arnold. Entretanto, é fácil perceber que ela se refere ao uso de Inteligência Artificial no processamento de linguagem natural e em visão computacional (termos mais correntes atualmente, apesar de NLU e IU ainda serem mencionados frequentemente). O artigo de Arnold é exemplar em vários pontos, inclusive por introduzir os semioticistas aos conceitos mais básicos do funcionamento de um computador. Entretanto, devido ao escopo do nosso trabalho, nos centraremos em apresentar o conceito de "representação do conhecimento" conforme proposto pela autora.

Redução e estruturação são noções-chave para a construção de sentido em semiótica; no entanto, eles não têm o mesmo propósito em IA. Na semiótica, a descrição estrutural é a última meta da análise, enquanto na IA é apenas um meio, imposto pela tecnologia do computador, de atingir os objetivos práticos da pesquisa. As diferenças entre as descrições elaboradas nos dois campos e o processamento ao qual eles dão origem podem ser explicadas em termos de seus objetivos e relações diferentes com a ciência da computação. (ARNOLD; 1988: 668)²²

Logo na introdução, a pesquisadora aponta que apesar de o ponto de vista sistematológico da semiótica ser um fim para esta ciência, ele pode ser um meio para a inteligência artificial, mesmo com os objetivos das duas áreas sendo ditos como diferentes. Para iniciar sua teoria, Arnold apresenta no artigo uma seção intitulada "A necessidade de explicitação e coerência do conhecimento", que inicia-se assim: "O fato de um computador funcionar por meio da realização de correspondências e por meio de álgebra booleana é a razão pela qual todo conhecimento fornecido a ele deve ser totalmente explícito e coerente." (ARNOLD, 1988: 671)²³. Ou seja, como dizíamos anteriormente, todo o conhecimento de um computador é fornecido pelo analista. Sem isso a máquina não é capaz de realizar absolutamente nada, como se o computador fosse 100% burro.

Considerando que a formalização do conhecimento, ou pelo menos sua explicitação e coerência rigorosa, é indispensável na IA, este não é o caso na semiótica por

²² Tradução nossa para: "Reduction and structuring are key notions for the construction of signification in semiotics; however, they do not serve the same purpose in AI. In semiotics, structural description is the ultimate goal of analysis, while in AI it is merely a means, imposed by computer technology, of attaining the practical goals of the research. The differences between the descriptions elaborated in the two fields and the processing to which they give rise can be explained in terms of their differing goals and relationships to computer science."

²³ Tradução nossa para: "The fact that a computer works by carrying out matchings and using Boolean algebra is the reason why all knowledge supplied to it must be totally explicit and coherent."

causa (1) das diferentes maneiras como o computador e o cérebro humano funcionam e (2) os diferentes objetivos perseguidos em ambas as disciplinas. Um ser humano, ao contrário de um computador (pelo menos até agora), é intuitivamente capaz de perceber formas, no sentido da palavra na psicologia da Gestalt (Dreyfus; 1972)²⁴. A estratégia de análise defendida por Hjelmslev (1968-71)²⁵ e aplicada sistematicamente na semiótica é por isso mesmo impossível atualmente para um sistema de computador, porque ele não pode de forma alguma chegar aos elementos primitivos por segmentações sucessivas; deve sempre partir dos constituintes mínimos de um texto. Além disso, uma vez que o objetivo da semiótica não é produzir sistemas capazes de realizar desempenhos cognitivos precisos, mas sim dar conta do fenômeno da significação em qualquer forma que ocorra, não é necessário para o corpus de conceitos, modelos e procedimentos que ela manipula ser imediatamente traduzível para uma linguagem formal ou ser totalmente explícito e coerente.²⁶ (ARNOLD; 1988: 676-677)

Neste excerto, Arnold sintetiza praticamente todos os problemas que discorreremos até agora. Em *Ciência da Programação*, o objetivo do pesquisador é desenvolver um sistema capaz de ser reproduzido em qualquer máquina e que realize alguma atividade com valor ao homem. Para isso, ele precisa formalizar totalmente a ação do computador para realizar tal objetivo. Diferentemente, a semiótica busca encontrar, subjacente aos textos, um sistema homogêneo capaz de explicá-lo por meio de sua reconstrução, mas não tem interesse algum em criar formalizações totalmente explícitas para que uma máquina possa aplicá-lo. Entretanto, o potencial existe, e pode ser explicitado por meio do caráter gerativo e não genético de nossa disciplina. No *Dicionário de Semiótica* encontramos o verbete Gerativo:

²⁴ Citação de Arnold: DREYFUS, H.; **What Computers Can't Do. The Limits of Artificial Intelligence**; New York: Harper and Row; 1972.

²⁵ Citação de Arnold: HJELMSLEV, L.; **Prolégomènes à une théorie du langage. Suivi de: La structure fondamentale du langage**; Paris: Minuit; 1968-71. Em nossa bibliografia: HJELMSLEV, L. [1943] 2018.

²⁶ Tradução nossa para: "Whereas knowledge formalization, or at least its explicitation and rigorous coherence, is indispensable in AI, this is not the case in semiotics because of (1) the different ways the computer and the human brain work and (2) the differing goals pursued in both disciplines. A human being, unlike a computer (up to now at least), is intuitively capable of perceiving forms, in the Gestalt psychology sense of the word (Dreyfus 1972). The strategy of analysis advocated by Hjelmslev (1968-71) and systematically applied in semiotics is for that very reason impossible at present for a computer system because it cannot at all get to the primitive elements by successive segmentations; it must always start out from the minimal constituents of a text. Furthermore, since the object of semiotics is not to produce systems capable of carrying out precise cognitive performances, but rather to account for the phenomenon of signification in whatever form it occurs, it is not necessary for the corpus of concepts, models, and procedures it manipulates to be immediately translatable into formal language or for it to be totally explicit and coherent."

Designamos pela expressão percurso gerativo a economia geral de uma teoria semiótica (ou apenas linguística), vale dizer, a disposição de seus componentes uns com relação aos outros, e isso na perspectiva de geração, isto é, postulando que, podendo todo objeto semiótico ser definido segundo o modo de sua produção, os componentes que intervêm nesse processo se articulam uns com os outros de acordo com um “percurso” que vai do mais simples ao mais complexo, do mais abstrato ao mais concreto. (GREIMAS e COURTÉS [1979] 2016: 232)

O percurso gerativo do sentido é um simulacro metodológico que busca recriar as condições para a significação. Por meio do quadrado semiótico²⁷, localizado no nível mais profundo, as oposições fundamentais de um texto são evidenciadas. No exemplo do patinho feio (MATTE e LARA, 2008: np), poderíamos pensar que as oposições mínimas presentes no texto são, entre outras, a /diferença/ e a /semelhança/ e como o patinho passa de uma para outra. Em relação ao esquema narrativo canônico²⁸, um nível acima do anterior, temos a disposição das estruturas actanciais. Sujeito, objeto e destinador são as principais posições que estão por trás de qualquer história. Por fim, no nível discursivo, temos a materialização desses elementos abstratos: o patinho feio é um sujeito em disjunção com a sociedade por ser diferente; após inúmeros processos, ele percebe que participa de outra comunidade, a dos cisnes, e entra em conjunção com ela. Notacionalmente, a história poderia ser representada como $S1 \rightarrow (S2 \wedge O)$, que se lê como: sujeito operador realiza uma transformação na relação de junção (da disjunção à conjunção) entre sujeito de estado e objeto de valor. Essas notações estão disponíveis como formalização para o analista fazer uma análise, entretanto, aplicações computacionais demandam mais do que apenas isso. Segue Arnold:

Um problema importante da inteligência artificial é reduzir e estruturar as informações contidas nos dados fornecidos ao computador - imagens, sons, textos escritos, pressões, etc. - e estabelecer descrições formalizadas capazes de serem processadas logicamente. A modelagem de comportamentos finalizados ocupa um lugar importante na representação do conhecimento. Parece que, neste campo em

²⁷ Em 2.3 definiremos e explicaremos melhor o funcionamento do quadrado semiótico quando falarmos sobre os diferentes tipos de sanção.

²⁸ Em 2.2 construiremos um modelo completo para representar o esquema narrativo canônico de um restaurante.

particular, o conhecimento narrativo e discursivo acumulado na semiótica deve dar uma contribuição significativa.²⁹ (ARNOLD; 1989: 81)

Todo um aparato metodológico e heurístico tem sido desenvolvido em mais de 50 anos de disciplina para produzir modelos capazes de “resumir” qualquer tipo de texto. Até mesmo uma receita de sopa foi minuciosamente analisada por meio da metalinguagem greimasiana. Nesse sentido, o trabalho de Madeleine Arnold é primoroso, pois vai, ponto a ponto, apresentando o estado da arte da Semiótica (seja por meio dos trabalhos de Greimas com língua natural ou mesmo os de Jean-Marie Floch, com a semiótica visual) e o da Inteligência Artificial mostrando os pontos em que se tocam e aqueles em que se afastam. Mais do que isso, Arnold se mantém sempre preocupada em aparar as arestas entre as duas ciências e propondo facilidades para que o semioticista aprenda a programar.

Outra grande diferença entre semiótica e IA deve ser enfatizada. Desta vez, ela não decorre do tipo de conhecimento processado, mas de restrições tecnológicas. Considerando que os modelos desenvolvidos em semiótica são válidos tanto em escala lexical quanto discursiva por força do princípio da elasticidade do discurso (Greimas e Courtés 1979: 86) sem a necessidade de dispositivos particulares na passagem de um nível a outro, este não é o caso em AI. Os modelos neste domínio devem incluir as palavras, meta-termos ou esquemas frasais que permitem que eles sejam iniciados quando aplicados em uma escala transfrástica. Desta forma, uma configuração discursiva, uma trajetória figurativa ou um papel temático podem ser reconhecidos mesmo na ausência de sua denominação correspondente.³⁰ (ARNOLD; 1988: 685)

Ainda hoje, quando se trata de processamento natural, as máquinas e os cientistas da computação encontram inúmeras dificuldades para se acercar da elasticidade da

²⁹ Tradução nossa para: “Un problème important en intelligence artificielle est de réduire et de structurer l'information contenue par les données fournies à l'ordinateur — images, sons, textes écrits, pressions, etc. — et d'en établir des descriptions formalisées aptes à être traitées logiquement. La modélisation des comportements finalisés occupe une place importante en représentation des connaissances. Il semble que dans ce domaine en particulier le savoir narratif et discursif accumulé en sémiotique devrait apporter une contribution non négligeable.”

³⁰ Tradução nossa para: “Another big difference between semiotics and AI must be emphasized. It stems this time not from the type of knowledge processed but from technological constraints. Whereas the models developed in semiotics are valid both on a lexical and a discursive scale by dint of the principle of discourse elasticity (Greimas and Courtés 1979: 86) without the need for particular devices when passing from one level to another, this is not the case in AI. The models in this domain must include the words, meta-terms, or phrase schemes which enable them to be initiated when applied on a transphrastic scale. In this way, a discursive configuration, a figurative trajectory, or a thematic role can be recognized even in the absence of its corresponding denomination.”

língua. O caso do patinho feio e das redações de vestibular são emblemáticos e iluminam-se mutuamente. Afinal, todos eles estão às voltas com o tema da aceitação social. “Chegar ao lago” ou “chegar em uma festa de desconhecidos”, semioticamente falando, parecem representar a mesma coisa nos casos apresentados. São variações sobre o tema do desconhecido e da coletividade. São formas de figurativizar fenômenos que possuem uma mesma base por baixo.³¹

Em processamento de linguagem natural, ferramentas de lematização até são capazes de verter /amaríamos/, /amou/ e /amava-se/ para um núcleo comum /amar/, o que ajuda bastante na hora de realizar análises quantitativas de corpus. Entretanto, em um texto bíblico, por exemplo, em que o termo Deus pode ser substituído por Ele, d’Ele, Sua, entre outras palavras com letra em maiúscula, os seres humanos dificilmente ficam sem entender a quem a partícula está se referindo. Porém, em computação, seria preciso criar regras, dicionários em que aquelas palavras são correlacionadas com o termo Deus³².

A elasticidade do discurso, ou seja, a capacidade de ele transcrever o mesmo conceito de forma condensada ou, ao contrário, expandida, impõe outro problema importante. Assim, a configuração discursiva da raiva, o percurso figurativo da pesca ou o papel temático do pescador podem aparecer em um texto verbal na forma condensada do lexema que os nomeia ou na forma ampliada de um enunciado transfrástico mais ou menos longo que, eventualmente, não incluirá as palavras “raiva”, “peixe” ou “pescador”.³³ (ARNOLD; 1989: 86)

Para Madeleine Arnold, entretanto, esses problemas existem, mas não devem assustar o semioticista. Muito pelo contrário. Segundo a pesquisadora, é exatamente

³¹ Em 2.1, quando apresentarmos nosso objeto de pesquisa, utilizaremos inúmeras tabelas que apresentam as sequências de palavras (n-gramas) mais frequentes nas avaliações de restaurante. Com isso, ficará claro que, quando quer referir-se ao serviço de um estabelecimento, um cliente pode utilizar inúmeras combinações possíveis para se acercar do mesmo fenômeno. Entretanto, isso ainda hoje é um problema no processamento de linguagem natural.

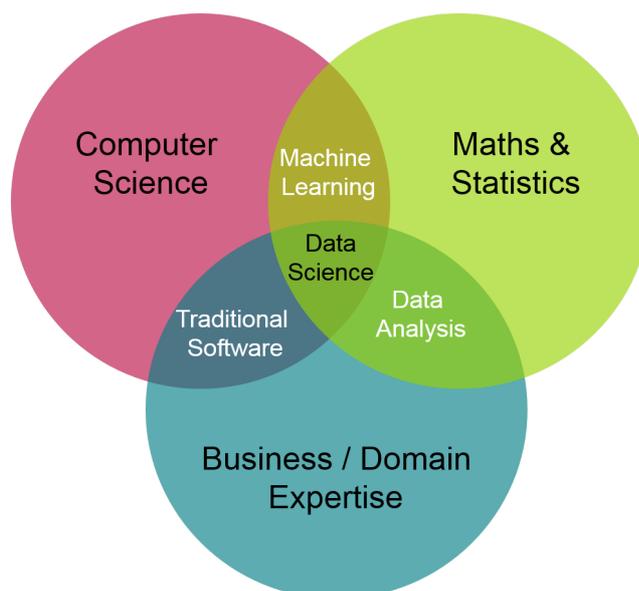
³² O mesmo problema já mencionado acima a respeito dos catafóricos ‘esse’, ‘este’ e ‘aquele’, mencionado anteriormente.

³³ Tradução nossa para: “L'élasticité du discours, c'est-à-dire la capacité de celui-ci à transcrire une même notion sous une forme condensée ou au contraire expansée, pose un autre problème important. Ainsi, la configuration discursive de la colère, le parcours figuratif de la pêche ou le rôle thématique du pêcheur peuvent apparaître dans un texte verbal sous la forme condensée du lexème qui les dénomme ou bien sous la forme expansée d'un énoncé transphrastique plus ou moins long qui, éventuellement, ne comportera pas les mots « colère », « pêcher » ou « pêcheur »”

para resolver esse tipo de problema que a semiótica se apresenta como uma ferramenta sem igual.

A semiótica dispõe de uma teoria global da significação, que falta atualmente na inteligência artificial. Assim, em semiótica, a análise de um texto, seja ele verbal, icônico ou outro - beneficia-se de uma economia geral que permite a junção de certos conhecimentos e a distribuição de outros de natureza mais específica, bem como *feedback* de análises específicas sobre o aparato teórico. É claro que, como enfatizamos até agora, as conquistas da semiótica geralmente não apresentam um grau suficiente de explicação e coerência para serem transferidas sem adaptações para os sistemas de inteligência artificial. No entanto, os fundamentos lógicos, a existência de procedimentos analíticos e a orientação formal da conceituação em semiótica são um bom presságio para as possibilidades de seu uso na inteligência artificial. ³⁴ (ARNOLD; 1989: 87)

Figura 4 - Diagrama de Venn da Ciência de Dados



Fonte: reprodução The Data Scientist³⁵

³⁴ Tradução nossa para: “La sémiotique dispose d'une théorie globale de la signification, ce dont l'intelligence artificielle est privée actuellement. Ainsi, en sémiotique, l'analyse d'un texte quel qu'il soit — verbal, iconique ou autre — profite d'une économie générale qui autorise la mise en commun de certaines connaissances et la distribution d'autres à caractère plus spécifique, ainsi que les rétroactions des analyses particulières sur l'appareil théorique. Bien sûr, comme nous l'avons souligné jusqu'ici, les acquis de la sémiotique ne présentent pas en général un degré d'explicitation et de cohérence suffisants pour qu'ils puissent être transférés sans adaptations dans des systèmes d'intelligence artificielle. Cependant, les fondements logiques, l'existence de procédures d'analyse et l'orientation formelle de la conceptualisation en sémiotique laissent bien augurer des possibilités d'une utilisation de celle-ci en intelligence artificielle.”

³⁵ Disponível em: <https://thedata scientist.com/data-science-without-programming/>. Consulta: 23/01/2021

Em Ciência de Dados (*Data Science*), existe um diagrama bastante conhecido (figura 4). Ele ilustra muito pertinentemente o fato de que de nada adianta você ter uma altíssima especialidade em Ciência da Computação e em Matemática e Estatística, se você não possuir algum tipo de conhecimento no domínio produtor dos dados ou no modelo de negócios que se interessa pelos dados. *Grosso modo*, sem uma orientação que signifique os dados para o ser humano, análises não parecem produzir resultados. Coisa que já era alertada por Charles Francis Hockett, estruturalista da Cornell, em 1952 nos Estados Unidos.

Não permitimos que o analista faça, simplesmente, jogos matemáticos com seus dados. Exigimos que ele produza sistematizações que, em um sentido operacional, correspondam aos hábitos que atribuímos ao falante: assim como o falante pode produzir qualquer número de novos enunciados a partir, essencialmente, do mesmo conjunto de hábitos subjacentes, a descrição do analista deve ser capaz de produzir qualquer número de novos enunciados, cada um capaz de passar no teste de aceitação casual por um falante nativo.³⁶ (HOCKETT; 1952: 98)

Ou seja, mais do que utilizar metodologias matemáticas a torto e direito, o analista precisa, antes de mais nada, organizar seus dados da forma mais coerente possível. Só assim a análise é capaz de produzir algum sentido “humano”. De certa forma, podemos resumir dizendo ser preferível que uma inteligência artificial seja capaz de produzir enunciados com sentido, não qualquer tipo de enunciado (como: dsfdfsd erwegds covfefefe rfsxrs) que não tem aderência alguma na nossa comunidade.

Acreditamos, portanto, que para Madeleine Arnold a principal contribuição da semiótica para a Inteligência Artificial é a capacidade de representação do conhecimento daquele grupo, no diagrama de Veen, conhecido como experiência de domínio ou negócios.

Duas maneiras de proceder estão, de fato, abertas ao semiótico: ou ele escolhe sistemas cujos modos de raciocínio e expressão são acessíveis ao usuário e o mais próximos possível aos dos seres humanos³⁷, ou ele trabalha com formalismos que

³⁶ Tradução nossa para Hockett, Ch. F.; "We do not allow the analyst simply to play mathematical games with his data. We require him to produce systematization which in an operational sense matches the habits which we ascribe to the speaker: just as the speaker can produce any number of new utterances from essentially the same set of underlying habits, so the analyst's description must be capable of producing any number of new utterances, each capable of passing the test of casual acceptance by a native speaker"

³⁷ Em 3.1, apresentamos a análise de sentimento como parte desta maneira de proceder.

podem não ter nada a ver com as atividades cognitivas humanas e que permanecem ocultos do usuário. Ambas as opções provaram ser eficazes. Linguistas e psicólogos, em particular, têm contribuído muito para a definição dos formalismos e para a implantação destes nos sistemas.³⁸ (ARNOLD; 1989: 86)

Nos próximos capítulos, portanto, iremos demonstrar o poder heurístico da semiótica por meio de exemplos que além de ilustrar todo o posicionamento de Arnold, poderão servir como atualização para artigos que foram escritos ainda na década de 1980, quando as técnicas de Inteligência artificial eram bastante incipientes.

³⁸ Tradução nossa para: “Deux façons de procéder s'offrent en effet au sémioticien: ou bien il choisit des systèmes dont les modes de raisonnement et d'expression sont accessibles à l'utilisateur et les plus proches possible de ceux d'êtres humains, ou bien il travaille sur des formalismes qui n'ont peut-être rien à voir avec les activités cognitives de l'homme et qui restent cachés à l'utilisateur. L'une comme l'autre voie ont démontré leur efficacité. Linguistes et psychologues, en particulier, ont fortement contribué à la définition de formalismes et à l'implantation de ceux-ci dans des systèmes.”

2. MODELO PRESSUPOSTO DA AVALIAÇÃO DE RESTAURANTE

Neste capítulo pretendemos construir o modelo pressuposto pela simples existência da avaliação de restaurantes no TripAdvisor. Nosso objetivo, portanto, será o de manter a coerência com a epistemologia semiótica, sintetizável na famosa frase de Greimas: "fora do texto não há salvação."

A epítome, pelo que parece, causou mais mal do que bem. Entretanto, a maioria dos críticos que enxergam na frase uma clausura, não compreendeu o pensamento greimasiano e muito menos leu algumas análises produzidas pelo pesquisador³⁹. Fato muito bem evidenciado e melhor combatido por Diana Luz na exposição de inúmeras formas de se cercar do "exterior" sem cair em "danação".

Muitas das críticas que a teoria recebeu e continua a sofrer são, assim, as de que se trata de uma proposta reducionista e a-historicista, voltada apenas para o exame da estruturação dos textos e discursos. São críticas que só se justificam pelo desconhecimento da teoria e dos rumos por ela assumidos em seus quase 50 anos. (BARROS; 2009: 352)

Quando propôs o famigerado *chavão*, Greimas deixou bastante claro que era possível ultrapassá-lo, sim, mas que haveria um custo nisso. Para o fundador da Escola de Paris, sempre houve a preocupação de que os semioticistas não caíssem em psicologismos e outras extravagâncias como: "Quando vocês encontrarem artigos sobre o sujeito da enunciação que pretendem dizer que o verdadeiro sujeito da enunciação é o pênis, certamente então sabemos onde isso vai dar." (GREIMAS; [1974] Sem Data: 25) Ou seja, a grande missão de Greimas com a epítome sempre foi a de que os pesquisadores não caíssem em equívocos de produzir correlações não pertinentes. Nos dias atuais, talvez pudéssemos atualizar a frase para algo como "fora do corpus não há salvação" ou "fora do *dataset*, não há salvação". Nesse sentido também evitaríamos estapafúrdias como a mencionada acima ou mesmo a de produzir asserções fortes a partir das correlações de dados que produzimos. Afinal, em ciência de dados precisamos sempre ter em mente que correlação não é relação de causalidade.

³⁹ Na seção 2.2 iremos apresentar brevemente partes de uma análise realizada por Algirdas J. Greimas a partir de uma receita de cozinha para evidenciar a metodologia conhecida como "classe e variação".

(...)há limites que não devem ser ultrapassados nesse sentido. E por quê? Por que podemos ultrapassá-los, mas é preciso avaliar o preço que vamos pagar se avançarmos. Isto quer dizer que tudo é possível para as pessoas que passam de um a outro, mas é preciso que o homem seja lúcido naquilo que faz e que não escorregue imperceptivelmente, que a vida seja um projeto voluntário e não um jogo de circunstâncias e deslizos cujo peso não se tenha avaliado de antemão. Vejam o que quis dizer: tomem, assumam uma direção ou outra, mas não por deslizos, por ignorância, por falta de lucidez. Aí está o verdadeiro problema da Semiótica. Portanto, a última palavra é essa, já disse antes: fora do texto não há salvação. Todo o texto, nada senão o texto e nada fora do texto. (GREIMAS; [1974] Sem Data: 25)

Sendo assim, na primeira seção faremos uma breve apresentação de nosso objeto: a avaliação de restaurante realizada na plataforma TripAdvisor. Lá, não discutiremos semiótica, mas apenas descreveremos o contexto histórico e econômico em que tais textos aparecem. Já na segunda seção, construiremos o esquema narrativo canônico daqueles textos avaliativos e proporemos abordar o ator da enunciação "usuário do TripAdvisor" como um sincretismo entre sujeito de estado (S2) e destinador-sancionador.

Por fim, a terceira seção se acercará da sanção propriamente dita. Para tanto, partiremos do quadrado semiótico clássico, em que figuram as sanções /positiva/, /negativa/, /não-positiva/ e /não-negativa/. Entretanto, será proposta uma discussão para atualização do dispositivo por meio da semiótica tensiva de Claude Zilberberg e do modelo catenário de Waldir Beividas (2019; e no prelo).

2.1. Objeto “avaliação de restaurante no TripAdvisor”

O TripAdvisor é uma das principais plataformas digitais de avaliação disponíveis atualmente. Por lá, hotéis, bares e restaurantes são avaliados por usuários de todo o planeta. Recentemente, entretanto, novas categorias como atrações turísticas, locadoras de veículos, companhias aéreas, cruzeiros, entre outras, começaram a ser disponibilizadas pela empresa americana.

Segundo divulgação do próprio TripAdvisor⁴⁰, eles possuem perto de 900 milhões de avaliações sobre quase nove milhões de locais. De acordo com informações do

⁴⁰ Disponível em: <https://tripadvisor.mediaroom.com/US-about-us>. Consulta: 20/11/2020

SimilarWeb, plataforma que fornece dados sobre o tráfego de usuários na internet, o site de avaliações se encontra em primeiro lugar entre os mais acessados na categoria turismo e viagem do mundo. No geral, TripAdvisor ocupa a 115ª posição entre os sites mais visitados dos Estados Unidos e a 335ª internacionalmente⁴¹. Entretanto, o principal concorrente da marca, que é o Yelp, aparece em outra categoria, o que inviabiliza o comparativo direto. Ainda assim, ele ocupa a 65ª posição entre os sites mais acessados dos EUA e a 222ª posição no mundo⁴² - o que, de certa forma, o coloca na frente da empresa foco de nossa pesquisa.

Porém, quando escolhemos o TripAdvisor como nosso objeto de pesquisa, e não o Yelp, levamos também outros fatores em consideração. Nossa meta sempre foi trabalhar com textos produzidos em português, daí priorizar os dados de circulação de tráfego das versões regionais de cada uma dessas plataformas. No Brasil, em questão numérica, TripAdvisor.com.br ocupa a 131ª⁴³ posição entre os sites mais acessados do país (o domínio usp.br, apenas como referência, ocupa a 99ª posição⁴⁴). Em contrapartida, por aqui, o Yelp fica com a 21.912ª, o que demonstra a baixa aderência à plataforma no Brasil. Outro fator decisivo na escolha da plataforma TripAdvisor para construirmos nosso corpus é a facilidade e a disponibilidade de *scripts* (códigos computacionais) para realizar a raspagem de dados das avaliações. Sendo assim, obter grandes quantidades de textos em português nunca foi um problema.

Agora, apesar de o fenômeno do TripAdvisor não ser novo, o valor que lhe é dado por chefs de cozinha e demais empresários do setor de restauração começou a crescer apenas recentemente. Até então, a principal preocupação do setor de alimentação fora do lar era com os críticos gastronômicos. Não por acaso, inúmeras reportagens dão conta de casos paroxísticos em que chefs levavam as avaliações da crítica especializada tão a sério, que tiravam a própria vida quando recebiam algum tipo de rebaixamento.

⁴¹ Disponível em: <https://www.similarweb.com/pt/website/tripadvisor.com/>. Consulta: 20/11/2020

⁴² Disponível em: <https://www.similarweb.com/pt/website/yelp.com/>. Consulta: 20/11/2020

⁴³ Disponível em: <https://www.similarweb.com/pt/website/tripadvisor.com.br/>. Consulta: 20/11/2020

⁴⁴ Disponível em: <https://www.similarweb.com/pt/website/usp.br/>. Consulta: 20/11/2020

Em 2003, o jornal *O Estado de São Paulo* escrevia que "Um dos mais famosos chefs da França, Bernard Loiseau, proprietário do Côte d'Or, em Saulieu, na Borgonha, um restaurante três-estrelas do *Guia Michelin*, matou-se com um tiro de fuzil de caça, após ter retrocedido na classificação de um guia concorrente, o *GaultMillau*."⁴⁵ Já em reportagem de 2019 da revista *Veja*, o chef Marc Veyrat, também da França, relata: "Sou o único chef na história que conseguiu uma terceira estrela em um ano e a perdeu no seguinte. Você pode imaginar a vergonha que isso representa? Estou em depressão desde que recebi a notícia"⁴⁶.

São apenas dois casos que encontramos em uma busca rápida na internet. Contudo, pesquisas mais aprofundadas fornecem uma verdadeira riqueza de fontes para descrever tal fenômeno que vitimizou não poucos cozinheiros. Não por acaso, é surpreendente quando um chef diz não se importar mais com estrelas. Caso de Daniel Boulud, badalado chef dos Estados Unidos que comanda mais de 15 casas. Em reportagem concedida à *Veja Rio*, ele diz que a preocupação mudou. "Meu restaurante, o Daniel, em Nova Iorque, tem duas estrelas, mas é o número 1 do TripAdvisor. E o TripAdvisor é fruto de uma avaliação do público. O Michelin não leva a opinião do público em consideração. Então, de certa maneira, isso me orgulha muito mais."⁴⁷ Em 2015, Daniel perdeu uma das três estrelas que seu restaurante principal detinha. Atualmente, ele é dono de uma das 13 casas nova-iorquinas que possuem 2 estrelas. Na mesma cidade, apenas 5 restaurantes possuem 3 estrelas.

Em se tratando de semiótica, não poucas pesquisas serão necessárias para dar conta dessa mudança apenas 'prenunciada', pois o fenômeno como um todo é muito mais complexo do que essas linhas introdutórias são capazes de descrever. Entretanto, gostaríamos de avançar uma pequena hipótese que leva em conta os manuais de marketing - no lugar dos elementos de análise do discurso.

⁴⁵ Disponível em: <https://internacional.estadao.com.br/noticias/geral,rebaixamento-em-guia-gastronomico-leva-chef-frances-a-morte,20030225p25786>. Consulta: 23/01/2021

⁴⁶ Disponível em: <https://veja.abril.com.br/gastronomia/processos-polemicas-e-suicidios-o-drama-das-estrelas-michelin/>. Consulta: 23/01/2021

⁴⁷ Disponível em: <https://vejario.abril.com.br/beira-mar/daniel-boulud-tripadvisor-michelin/>. Consulta: 23/01/2021

“Guru do marketing”⁴⁸, Philip Kotler explica em Marketing 4.0 o fenômeno do TripAdvisor (e de outras plataformas digitais) propondo uma categoria de consumidores que define como “netizens ou cidadãos da internet”. Para o professor americano, junto às categorias de mulheres e jovens, esses consumidores são responsáveis pela “defesa da marca no mundo digital”. São populações de indivíduos que normalmente se engajam na produção de avaliações e na disseminação de informações sobre empresas das quais se convertem em verdadeiros advogados.

Para aumentar a probabilidade de conquistar defensores da marca, os profissionais de marketing deveriam apostar nos JMN: jovens, mulheres e netizens ou cidadãos da internet. Muitos temas relacionados a esses três grandes segmentos foram pesquisados e explorados separadamente. Em termos de tamanho, cada um deles representa uma parte bastante lucrativa. Portanto, a abordagem de marketing tem sido feita sob medida para atendê-los. Só que existe um quadro mais amplo. Há um fio condutor comum que faz uma ligação entre eles: os JMN são os segmentos mais influentes na era digital. (KOTLER; 2017: 48)

Tal fato não é desconhecido de grande parte dos empresários de alimentação fora do lar. Tanto é que Daniel Boulud, na mesma entrevista em que menciona se importar mais com o TripAdvisor, hoje em dia, do que com as estrelas dos guias especializados, define bastante bem a situação. “O poder das mídias sociais está nas mãos de quem sabe usá-las.” Para definir esse bom uso, é necessário compreender melhor quais são as mãos poderosas por trás das redes sociais digitais, algo que Kotler faz bastante bem.

Existem 3,4 bilhões de usuários da internet - 45% da população mundial, de acordo com estimativas das Nações Unidas. Nem todos podem ser considerados netizens ou cidadãos da internet - e a segmentação da Forrester's Social Technographics pode ajudar a explicar por quê. De acordo com essa segmentação, existe uma hierarquia de usuários da internet, que inclui inativos, espectadores (pessoas que leem e assistem a conteúdo on-line), participantes (pessoas que mantêm e visitam páginas de mídias sociais), coletores (pessoas que acrescentam *tags* às páginas da *web* e usam feeds RSS), críticos (pessoas que postam avaliações e comentários on-line) e criadores (pessoas que criam e publicam conteúdo on-line). Os coletores,

⁴⁸ A expressão já se tornou um verdadeiro clichê para se referir ao pesquisador. Em <https://exame.com/revista-exame/licoes-para-o-pos-covid/> é possível conferir o emprego dela em uma reportagem da revista *Exame*. Consulta: 23/01/2021

críticos e criadores caracterizam melhor os netizens - indivíduos que ativamente contribuem para a internet, não se limitando a consumir a internet. (KOTLER; 2017: 55)

Nossos atores da enunciação (conceito semiótico que será melhor descrito na seção 2.2) encontram-se muito bem inseridos na citação acima. Eles são os críticos, os cidadãos que dedicam um tempo de sua vida a postar avaliações no ambiente online. Para Kotler, mais do que o papel ativo no desenvolvimento da internet como um espaço horizontalizado de democracia total, tais sujeitos possuem o poder de convencimento e de mobilização de todas as outras classes. Afinal, mantendo nosso objeto em mente, não são poucos os usuários do TripAdvisor que se caracterizam apenas como espectadores. Pessoas que, antes de realizar a reserva de um hotel ou de planejar o jantar durante uma viagem, sempre recorrem àquela plataforma para tomar suas decisões - inclusive por não haver a necessidade de se registrar para consultar a produção dos usuários cadastrados.

O papel deles de influenciar os outros está ligado ao desejo de estar sempre conectado e contribuir. Os netizens são conectores sociais. Sabemos que os netizens adoram se conectar. (...)Os netizens, ao contrário dos usuários da internet em geral, apresentam maior tendência a se tornarem advogados de marcas. (...)Além disso, os advogados de marcas são também contadores de histórias que espalham notícias sobre as marcas pelas suas redes. Eles contam histórias autênticas do ponto de vista do cliente - um papel que a publicidade nunca poderá substituir. (...)São chamados de cidadãos da internet por um motivo: à semelhança de bons cidadãos que contribuem para seu país, eles contribuem para o desenvolvimento da internet. (...)Com avaliações e análises de produtos na internet, outras pessoas podem facilmente descobrir a melhor opção possível. (KOTLER; 2017: 55-57)

Como dissemos desde o começo, nosso interesse nessas poucas linhas introdutórias é apenas a de prenunciar uma problemática bastante maior de um fenômeno ainda pouco conhecido - pelo menos do ponto de vista semiótico. Propor a utilização dos manuais de marketing, no lugar dos elementos de análise do discurso, para dar uma rápida pincelada no assunto, tem como principal fator demonstrar que há uma valorização crescente entre os profissionais da propaganda no endosso da conversão de todo e qualquer cliente em um netizen. É economicamente bastante lucrativo (e barato) incentivar cada vez mais pessoas a

avaliar os restaurantes que frequentam. Não por acaso, são inúmeros os mimos para que façam isso. Em se tratando do setor de alimentação fora do lar, então, muitas são as cortesias oferecidas aos clientes que realizam uma avaliação: drinques gratuitos, couvert não cobrado, descontos etc.

Sendo assim, para iniciar qualquer pesquisa que queira entender melhor o fenômeno brevemente exposto, é premente que uma metodologia de análise do discurso computadorizada seja desenvolvida. Sem uma maneira de sintetizar todas as avaliações, uma forma de quantificar semanticamente os valores expressados pelos usuários naquela plataforma, nos parece que qualquer pesquisa já parte com uma desvantagem bastante grande. Portanto, nosso objetivo é reduzir esse déficit metodológico oferecendo uma primeira aproximação entre semiótica discursiva e inteligência artificial.

Antes disso, entretanto, algumas estratégias de análise quantitativa de corpus, já bem conhecidas da linguística computacional, ajudam a entender melhor os nossos textos. Entre elas, a extração de n-gramas parece ser bastante útil para o semioticista, uma vez que é capaz de fazer ver as principais figuras discursivas presentes em um corpus. Segundo FERREIRA e LOPES (2019: 129), “Um n-grama é simplesmente uma sequência de n elementos em um determinado nível de análise (letras, morfemas, palavras etc.). (...)A depender do valor de n , falamos em unigramas, bigramas, trigramas, quadrigramas...”

Abaixo, apresentamos os 10 primeiros termos de quatro formações de n-gramas diferentes (bigramas, trigramas, quadrigramas e pentagramas) do nosso corpus bruto, formado por 45 mil avaliações⁴⁹.

Tabela 1 - Os 10 bigramas mais frequentes

Sequência de palavras (2-gramas)	Frequência
O atendimento	5814
Muito bem	5521
E o	5466
Muito bom	5152
A comida	4493

⁴⁹ Mais informações sobre o corpus estão disponíveis em 3.3 e no apêndice

Atendimento do	3623
Ótimo atendimento	3373
Excelente atendimento	3036
E a	2989
Os pratos	2480

Fonte: produção própria

Tabela 2 - Os 10 trigramas mais frequentes

Sequência de palavras (3-gramas)	Frequência
Fomos muito bem	1466
Vale a pena	1387
A comida é	1342
O atendimento do	1241
Muito bem atendidos	1180
E o atendimento	1174
Atendimento do garçom	1112
O atendimento é	1011
Fomos atendidos pelo	963
Frutos do mar	944

Fonte: produção própria

Tabela 3 - Os 10 quadrigramas mais frequentes

Sequência de palavras (4-gramas)	Frequência
Fomos muito bem atendidos	995
Muito bem atendidos pelo	655
Vale muito a pena	481
Nos atendeu muito bem	396
De frutos do mar	351
O Atendimento do garçom	350
Fomos atendidos pelo garçom	289
E fomos muito bem	288

Bem atendidos pelo garçom	269
E o atendimento do	265

Fonte: produção própria

Tabela 4 - Os 10 pentagramas mais frequentes

Sequência de palavras (5-gramas)	Frequência
Fomos muito bem atendidos pelo	562
Muito bem atendidos pelo garçom	220
E fomos muito bem atendidos	201
A comida é muito boa	156
O atendimento é muito bom	96
Vale a pena a visita	92
O ambiente é muito agradável	90
Os pratos são bem servidos	86
Comemorar o aniversário da minha	75
Cocada ao forno com sorvete	75

Fonte: produção própria

Conforme dizíamos, essa é apenas uma primeira forma de apresentar quantitativamente o nosso objeto. Aparentemente, de acordo com o levantamento, figuras ligadas ao serviço são mais abundantes do que aquelas que expressam algum tipo de asserção sobre variedades de alimentos ou características arquitetônicas e decorativas do estabelecimento.

Esse tipo de análise, todavia, não é novidade em semiótica. Em 1987, por exemplo, Jean-Marie Floch publica artigo em que demonstra a utilidade de nossa disciplina na concepção de um espaço comercial. Para realizar essa empreitada, o pesquisador entrevistou inúmeros consumidores em busca de descobrir o que seria um supermercado ideal. A partir desses dados, ele desenhou a arquitetura de um espaço que dava conta de todos os anseios dos consumidores. As principais expressões utilizadas pelos entrevistados podem ser vistas na figura 5.

A partir de tais partículas discursivas, Floch foi capaz de criar uma tipologia de consumidores por meio de um quadrado semiótico em que encontravam-se os valores /práticos/, /utópicos/, /críticos/ e /lúdicos/. Por baixo de cada entrevista concedida pelos compradores de supermercado, o pesquisador foi capaz de encontrar esses valores mínimos.

Figura 5 - O hipermercado ideal. Principais características mencionadas por 400 consumidores entrevistados pela Domas_Ipsos em junho de 1986.

O hipermercado ideal como hipermercado:	Essencial	Ranking				Média do Ranking
		1º	2º	3º	Outros	
É Iluminado, limpo e organizado	74%	19%	31%	13%	11%	2.2
Onde tudo é feito no menor tempo possível	97%	63%	21%	8%	5%	1.5
Onde eu posso encontrar a informação que eu preciso para poder escolher	67%	6%	20%	26%	15%	2.8
Onde eu posso passear com prazer	45%	4%	7%	9%	25%	3.5
Que tem personalidade, alma	42%	2%	10%	12%	18%	3.4
Onde eu posso compartilhar as compras	28%	1%	2%	9%	16%	3.7
Onde eu encontro um lugar para relaxar	26%	3%	5%	7%	11%	3.4
Onde eu posso encontrar pessoas, como nos mercados	15%	1%	2%	2%	10%	4.0

Fonte: reprodução (FLOCH; [1987] 2014: 37)

esteja o discurso em sua forma expandida (longas descrições, narrativas detalhadas) ou em forma concisa (um adjetivo, um verbo, uma comparação breve), a análise semiótica se faz a partir do reconhecimento das várias formas de programas narrativos e dos “objetos de valor” que são neles recorrentes e os definem, um em relação ao outro. (FLOCH; [1987] 2014: 28)

Para a nossa pesquisa, entretanto, o modelo flochiano não se apresenta muito frutífero, uma vez que a maioria dos n-gramas revela uma hipertrofia de valores /práticos/, aqueles que destacam o valor de uso do restaurante: serviço de qualidade, comida saborosa entre outros. Valores utópicos, ou míticos, relativos à saúde, ao meio ambiente, etc., não parecem ser muito importantes para o usuário do TripAdvisor. Dificilmente vemos menção à preservação do ambiente ou sustentabilidade nos discursos apresentados. Por outro lado, valores críticos estão razoavelmente presentes e utilizam bastante a figurativização “preço salgado”, por exemplo, para falar que a faixa de valores praticada em um estabelecimento é alta.

Agora, a alta densidade de expressões sobre o serviço também não é nenhuma novidade. Reportagem do jornal *O Globo* de 14 de julho de 2020 dá conta de tal fato já em seu título: “Comensal valoriza mais atendimento do que comida em restaurantes do Rio e de São Paulo, aponta pesquisa”.

Uma pesquisa do grupo Consumo, Gastronomia e Redes Sociais, da UFRJ, aponta o que preferem os clientes de restaurantes do Rio e em São Paulo. A partir da leitura estatística de 16 mil comentários publicados nos perfis dos dez melhores do TripAdvisor, os pesquisadores verificaram que o atendimento é o quesito mais citado pelos clientes, seguido da comida.⁵⁰

Os pesquisadores em questão também utilizaram modelos de n-gramas para construir as nuvens de palavras mais fortes nos restaurantes cariocas. "Quando recorremos ao software estatístico R para criar os bigramas ou trigramas (associação de palavras), vemos as correlações formadas: 'Atendimento Excelente', 'Ambiente super agradável' e 'comida maravilhosa'."⁵¹ No nosso caso, os modelos de n-gramas que construímos também davam conta de algo bastante semelhante. Infelizmente, até o presente momento da pesquisa, não obtivemos mais informações a respeito do trabalho realizado pelos pesquisadores da UFRJ. Consultas nas principais plataformas de publicação acadêmica pelo nome dos profissionais envolvidos não retornou nenhuma publicação que pudesse nos ajudar a entender melhor o fenômeno avaliação de restaurantes.

Figura 6 - O objeto “avaliação de restaurante no TripAdvisor” no destaque em vermelho. Os demais campos são considerados metatextos a serem correlacionados posteriormente.

Publicada em 29 de março de 2020 via dispositivo móvel

Requinte e sabor

O ambiente é muito requintado, o atendimento é impecável, adorei a harmonização. O menu optimus é algo que realmente desperta o sentido, os sabores são intrigantes e valorizam a experiência. Apesar de as porções serem mínimas eu saí com a sensação de ter comido em excesso.

Mostrar menos

Data da visita: fevereiro de 2020

[Ver todas as 57 avaliações de WelderMedeiros sobre São Paulo](#)

[Peça informações para WelderMedeiros sobre D.O.M. Gastronomia Brasileira.](#)

Esta avaliação representa a opinião subjetiva de um membro do programa TripAdvisor e não da TripAdvisor LLC.

Fonte: reprodução/ TripAdvisor

Finalmente, antes de terminarmos esta seção, é importante fazer uma formalização maior de nosso objeto. Portanto, apresentamos na figura 6 uma reprodução de

⁵⁰ Disponível em: <https://blogs.oglobo.globo.com/luciana-froes/post/comensal-valoriza-mais-atendimento-do-que-comida-em-restaurantes-do-rio-e-de-sao-paulo-aponta-pesquisa.html>. Consulta: 23/01/2021

⁵¹ *Idem*

avaliação conforme ela aparece após ter sido publicada na plataforma. Definimos como corpus de análise apenas o texto contido dentro do retângulo vermelho. Todas as demais informações (metatextos), como a imagem e o nome do usuário, bem como a localização dele durante a produção do discurso, a data de publicação etc. foram deixados de lado.

Acreditamos ser metodologicamente mais correto iniciar nossos trabalhos apenas pela análise (e produção de dados) dos textos a que chamamos de avaliação (em destaque na figura 6). Após essa etapa, outras análises poderão ser realizadas a partir da correlação entre os textos e os metatextos. Por exemplo, será possível verificar a possibilidade de inferência das classes propostas⁵² a partir das palavras do título, o espaço temporal médio entre data de publicação e de visita em relação à cada categoria (haverá alguma relação entre ascendência e descendência tensivas e urgência na produção de avaliação?), entre outras possibilidades.

2.2. Contexto situacional e estrutura actancial do enunciado

Se há uma avaliação/sanção, isso quer dizer que houve uma performance. Essa, também pressupostamente, exige a existência de uma manipulação que, anteriormente, instituiu um contrato, além da competencialização que forneceu um poder e um saber-fazer. Falando assim, as coisas parecem complicadas. Entretanto, por meio de exemplos, acreditamos que poderão ficar mais fáceis.

A partir da disposição dos eus, dos tus e dos eles, dos tempos verbais ou de figuras de espaço (aqui, ali, acolá etc.) somos capazes de encontrar um sistema narrativo por baixo de todos os enunciados com que nos deparamos. Alguns dão mais trabalho, como é o caso daqueles conhecidos como “texto em terceira pessoa”. Em metalinguagem semiótica, o tipo é descrito como enunciado enuncivo. Ou seja, o assunto fala-se por si mesmo. Na verdade, este é um efeito de sentido bastante estudado e faz parte das ilusões referenciais que buscam causar maior objetividade (em oposição à subjetividade). Todavia, até nesses textos, um “eu” está sempre pressuposto.

⁵² Recrudescimento, saturação, atenuação... Nossa classificação completa é apresentada em 3.2.1.

Pode-se dizer, por exemplo, “A Terra é redonda” ou “Eu digo que a Terra é redonda”. Isso significa que se pode afastar o “eu” do enunciado e ter-se-ia, então, aquilo que na Teoria Literária se chama uma narrativa em terceira pessoa; ou pode-se instalar o “eu” no interior do enunciado, pode-se projetar a própria enunciação dentro do enunciado e aí se tem uma enunciação enunciada, uma narrativa em primeira pessoa. Quando se diz: “Eu digo que a Terra é redonda”, projeta-se o ato de dizer no dito, diz-se a enunciação, enuncia-se o ato de dizer. (FIORIN; 2017: 977)

Ou seja, pela simples existência de um enunciado (seja ele enuncivo ou enunciativo), sempre se depreende uma enunciação. Até do simples preenchimento de dados em uma tabela pode ser encontrado, dentro de um contexto situacional, um esquema narrativo. Veja o que diz Greimas, por exemplo, a respeito de um dos mais corriqueiros dispositivos discursivos: uma receita de sopa.

A receita de cozinha, embora formulada, na superfície, com a ajuda de imperativos, não pode ser considerada como uma prescrição, regida por um /dever-fazer/ subentendido no conjunto do texto. Ela se apresenta antes como uma proposição de contrato do tipo: "Se você executar corretamente o conjunto de indicações dadas, você obterá a sopa ao 'pistou'". Trata-se, portanto, de saída, de uma estrutura actancial pondo em presença dois sujeitos - o destinador e o destinatário - situados ambos sobre a dimensão cognitiva: o destinador, normalmente encarregado de modalizar o destinatário, não se preocupa em lhe transmitir um /querer-fazer/ nem um /dever-fazer/, contentando-se em investi-lo da modalidade do /saber-fazer/. (...)A receita de cozinha pode, por conseguinte, ser considerada como uma subclasse de discursos que, analogamente às partituras musicais ou aos projetos do arquiteto, apresentam-se como manifestações de competência atualizada, anteriormente à sua realização. (GREIMAS; [1979] 1996: 10-11)

"Debulhe", "descasque", "faça cozinhar", "coloque sua sopa numa sopeira e a sirva" são manifestações discursivas que evidenciam um 'eu' e um 'tu' bastante genéricos mas sempre presentes nas receitas culinárias. São esses dois sujeitos, um destinador e um destinatário, que instituem o "estatuto modal do texto" e permitem que generalizemos todas as receitas como uma classe/subclasse de discursos. Voltando a *Semiótica e Ciências Sociais*, Greimas define melhor a metodologia/epistemologia utilizada para fazer tal tipo de generalização.

A ausência do contexto pode também ser suprida pelo recurso às variantes. Já sabemos um certo número de coisas sobre o mito e suas variações, sobre o conto e

suas variantes. Creio que é preciso reter a definição, já clássica, de Lévi-Strauss e dizer, generalizando, que a narrativa não é uma mensagem-ocorrência autônoma mas que é constituída pelo conjunto de correlações entre todas suas variantes. Esta definição, aliás, nada mais faz que assumir a concepção de Hjelmslev, para quem a gramática comparada das línguas indo-europeias, por exemplo, constitui apenas um sistema de correlações entre as línguas particulares através das quais ele se manifesta, concepção que remete, por sua vez, aos trabalhos de Saussure, indo-europeísta do fim do século XIX. A reunião do maior número possível de variantes, esclarecendo-se umas às outras, permitiria assim reconstituir o conteúdo investido na narrativa considerada enquanto classe. (GREIMAS; [1976] 1981: 173-174)

Nossa intenção nesta seção, portanto, é a de constituir uma classe de discursos por meio de textos recolhidos na plataforma TripAdvisor. Em todos eles, igual acontece com a receita, parece que temos também dois actantes frente a frente: destinador e destinatário. A primeira posição é ocupada pelo usuário que realiza uma avaliação, enquanto que a segunda fica à cargo do restaurante. Para esclarecer melhor esse ponto, consultaremos um dicionário para depreender as estruturas mínimas de um restaurante. Entretanto, antes disso, vale a pena explicitar tal prática, tão frequente na semiótica, que é a de consultar dicionários para obter as estruturas narrativas.

Colocou-se o mesmo problema a respeito dos provérbios. À minha objeção, a saber que sua pesquisa sobre os provérbios não levava em conta a distinção entre o significante e o significado, Jean Cuisenier respondeu, não sem razão, que ele não dispunha de processos objetivos que permitissem o reconhecimento dos significados. Minha resposta - e utilizo aqui as conversas de bastidores - foi que os significados, de fato, só podiam ser reconhecidos inscrevendo os provérbios em sua isotopia contextual. A contraproposta de Jean Cuisenier merece ser destacada: na falta do contexto que pressupõe o registro de cada provérbio captado no discurso "espontâneo", poder-se-ia conceber o recurso a uma espécie de Comitê dos Sábios, a um comitê de especialistas de uma determinada cultura, cujo consenso equivaleria à referência explícita ao código natural em questão. (GREIMAS; [1976] 1981: 173)

Para Greimas, de acordo com a citação acima, em que se fala sobre uma solução frástica recorrente na fala de vários informantes, os provérbios podem ser tomados como significante de um significado. Ou seja, uma partícula inteira ("deus te crie", depois de um espirro, "devagar com o andor que o santo é de barro", quando alguém apresenta uma atitude estabonada) só é plenamente realizada quando se conhece o contexto. E a proposta de Jean Cuisenier para resolver esse problema

não poderia ser mais exuberante: a criação de um "Comitê dos Sábios". Ora, isso não é exatamente o trabalho de um dicionarista? Não podemos pensar inclusive em dicionários de expressões e dicionários filosóficos? A elasticidade da língua sempre nos ronda.

Levando em consideração que nosso objeto institui uma sanção (mais especificamente, a avaliação de um restaurante), é preciso, antes de mais nada, definir a performance de tal sujeito. Ainda que em contextos diferentes, também naquele texto de 1970 Greimas apresenta a importância do uso dos dicionários para a definição do universo semântico em que uma análise se inscreve.

A presença - ainda parcial - do contexto situacional implícito deve conduzir-nos à constituição de um dicionário mitológico, isto é, ao estabelecimento, para cada cultura determinada, de um inventário lexemático feito de palavras ou de expressões que, no momento da análise concreta, parecem "significativos" pelo fato de que, por exemplo, manifestavam actantes ou funções dentro do esquema narrativo entrevisto. Tal dicionário, evidentemente, só deveria ser estabelecido para ser transformado, depois da análise, em código semântico de um determinado universo cultural. (GREIMAS; [1976] 1981: 171) ⁵³

Bastante densa, a citação acima se inscreve dentro de um contexto em que os pesquisadores que participaram do Simpósio de Literatura Técnica buscavam inferir os significados das narrativas em que já haviam delimitado os significantes. Tais termos aparecem no texto de Greimas em um sentido bastante mais ampliado do que nas tradicionais dicotomias saussureanas. As entidades narrativas (destinador, sujeito e objeto, por exemplo) ficam um tanto quanto empobrecidas quando não inscritas dentro de um universo semântico. No caso específico, a discussão do congresso girava em torno, principalmente, dos provérbios e, antes mesmo da proposta do dicionário, Jean Cuisenier, como apresentamos, propôs a concepção de um 'Comitê dos Sábios' para dinamizar melhor as estruturas narrativas: algo como fornecer o revestimento temático figurativo pertinente a um universo semântico particular.

Se esquecêssemos completamente o que sabemos sobre um restaurante, por exemplo, e lêssemos algo como "O restaurante Andorinha não me deixou saciado".

⁵³ Retomamos citação já apresentada em 1.2 para evitar remissões a pontos longínquos do texto.

Apesar do desconhecimento semântico, um semioticista ainda poderia, munido das estruturas analíticas, afirmar que um sujeito S1 não realizou a performance de transformação da relação de junção do sujeito S2 com a sociedade. E nada mais. Afinal, desconheceríamos as implicações que o lexema restaurante ocupa dentro do contexto sócio-cultural daquela frase. Tais afirmações parecem pouco pertinentes em pleno século XXI, quando os universos do sentido estão cada vez mais homogeneizados. Entretanto, em 1970, quando grande parte dos pesquisadores debruçava-se sobre relatos míticos (de povos indígenas) ou contos populares, fazia todo sentido questionar o revestimento temático-figurativo⁵⁴ que as posições actanciais assumiam dentro de uma cultura.

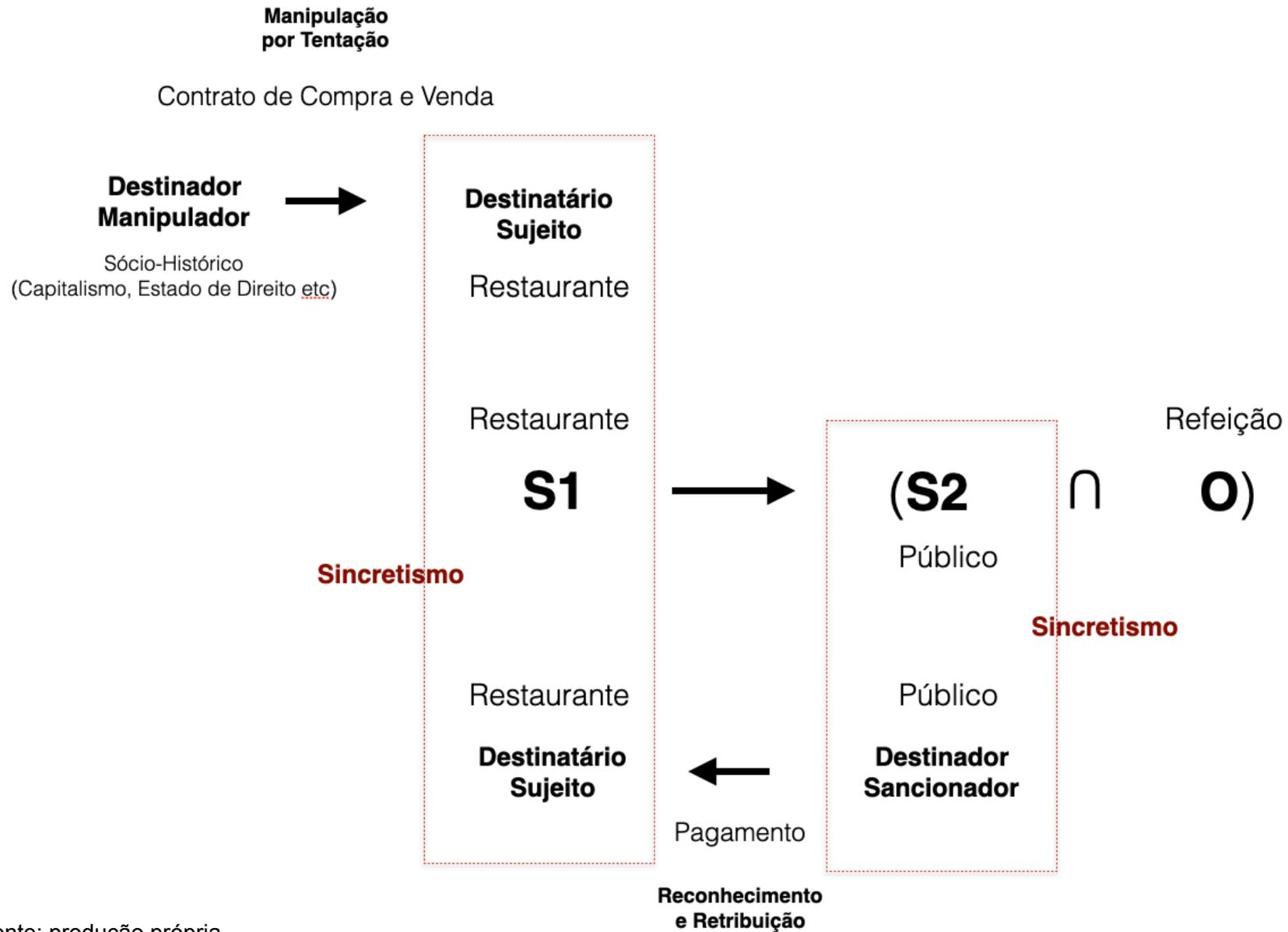
Felizmente, não precisamos constituir comitê algum. Basta uma consulta rápida ao dicionário Michaelis, por exemplo, para encontrarmos no verbete restaurante a seguinte definição: "Estabelecimento comercial onde se servem refeições ao público, mediante pagamento." A partir daí, conhecendo as posições actanciais e o esquema narrativo canônico, facilmente depreendemos o modelo narrativo canônico do restaurante (figura 7).

A primeira parte da definição do dicionário apresenta a performance e os actantes sujeito operador (S1), sujeito de estado (S2) e objeto (O). Ou seja, um restaurante é um sujeito que opera a transformação na relação de junção do sujeito público com o objeto refeição. Já na segunda parte, um contrato fiduciário é evidenciado quando a preposição 'mediante' entra em jogo. Como já havíamos mencionado a respeito do patinho feio (a partir da página 22 desta dissertação), a lógica da narrativa é diferente da lógica do discurso. Apesar de ser o último termo da definição do dicionário, o pagamento é um das primeiras relações de pressuposição recíprocas instituídas pela simples existência de um restaurante.

Conforme pode ser visto na figura 7, existe um destinador-manipulador sócio-histórico (sobre o qual nada pode ser dito sem que se envie a análise a favor de uma ou de outra concepção de política e/ou de economia). Esse actante, que muitos

⁵⁴ Fato já evidenciado e ressaltado por Pessoa de Barros ao dizer que a competência figurativa do falante é sócio-histórico. Confira trechos do debate em 1.2.

Figura 7 - Modelo narrativo canônico do restaurante (ponto de vista do pagamento). Em negrito estão as posições actanciais.



Fonte: produção própria

podem, apressadamente, caracterizar como o capitalismo, só pode ser definido como o sujeito que fornece um querer-fazer ou dever-fazer ao sujeito restaurante mediante uma manipulação por tentação e um contrato. "Se você servir uma refeição, você ganha dinheiro" acaba sendo, nos parece, a fórmula que melhor sintetiza tal situação⁵⁵ - e que também serve para inúmeros outros fenômenos. Apesar de não fazer parte do escopo deste trabalho, vale lembrar a tipologia das manipulações fornecida por Barros.

Há quatro grandes tipos de figuras da manipulação, a provocação, a sedução, a tentação e intimidação, segundo dois critérios de classificação: o da competência do manipulador para o fazer persuasivo e o da alteração modal operada na competência do sujeito manipulado. No primeiro caso, o destinador-manipulador persuade pelo saber, provocando e seduzindo, ou pelo poder, tentando e intimidando. Na provocação e na sedução, o destinador diz ao destinatário, de forma clara ou implícita, o que sabe de sua competência, colocando-o em posição de escolha forçada. Na provocação, deve escolher entre aceitar a imagem desfavorável que dele foi apresentada ou fazer o que o manipulador pretende; na sedução, precisa recusar a representação lisonjeira que dele foi feita ou deixar-se manipular. O julgamento da competência é, portanto, positivo, na sedução, e negativo, na provocação. Na tentação e na intimidação, o manipulador mostra poder e propõe ao manipulado, para que ele faça o esperado, objetos de valor cultural, respectivamente positivo (dinheiro, presentes, vantagens) e negativo (ameaças). (BARROS; 2002: 38)

Com isso, nos livramos de inúmeras inconveniências e definimos nossos termos apenas pela posição (relacional) que assumem entre si. Sem ontologia e sem metafísica. Sendo assim, agora estamos suficientemente munidos para falar sobre o "ator da enunciação".

Segundo o *Dicionário de Semiótica*, "Para ser chamado de ator um lexema deve ser portador de pelo menos um papel actancial e de no mínimo um papel temático." (GREIMAS e COURTÉS; [1979] 2016: 45) Por papel actancial entende-se, conforme explicitado acima, as funções de sujeito operador (S1), sujeito de estado (S2), objeto (O), destinador-manipulador, destinador-competencializador e destinador-sancionador. Enquanto que, por papel temático, entende-se a figura de

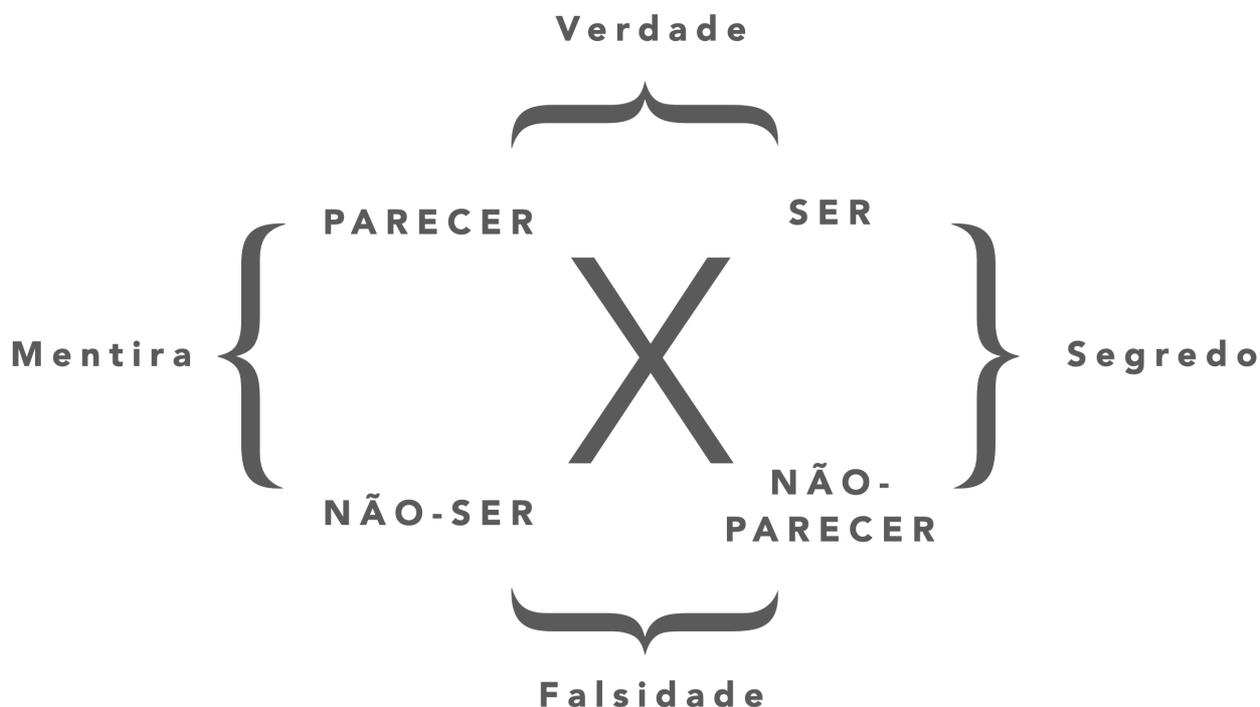
⁵⁵ É preciso notar que, neste caso, o público é um delegado do destinador sócio-histórico sobre o qual nada se pode falar sem enviar a análise para uma ou outra representação cultural da sociedade e da economia.

um "personagem" dentro de uma história: um pescador, por exemplo, e tudo o que se pode inferir a partir dele. Sendo assim, para se obter a figura do ator é preciso a união das duas coisas. Conforme visto na figura 7, o 'usuário do TripAdvisor' (que é um papel temático), parece ocupar duas casas: sujeito de estado, pois é ele que tem sua realidade transformada (ele entra em conjunção com a refeição), e destinador-sancionador (delegado), uma vez que se espera que ele faça o pagamento (em nome de um contrato pressuposto). Mais pesquisas precisam ser feitas para definir melhor o lugar deste ator em relação ao destinador-manipulador sócio-histórico⁵⁶.

2.3. Tipologia das sanções de restaurante

Não são poucas as pesquisas que trabalharam o conceito de sanção. Grande parte delas, entretanto, por muito tempo têm privilegiado a dimensão cognitiva desta estrutura narrativa. O quadrado da veridicção, não por acaso, ainda hoje conserva uma força de representação bastante forte. Dificilmente um estudante de semiótica desconhece a estrutura representada na figura 8. Apesar de o escopo de nosso trabalho não centrar nessa dimensão, cabem algumas palavras sobre essa diferenciação.

⁵⁶ Conforme evidenciamos na figura 7, estamos lidando apenas com um dos muitos pontos de vista possíveis. Mais pesquisas são necessárias para lidar com as demais manipulações, competencializações, performances e sanções que também estão em jogo.

Figura 8 - Quadrado da veridicção⁵⁷

Fonte: produção própria a partir do *Dicionário de Semiótica* (GREIMAS e COURTÉS; [1979] 2016: 532)

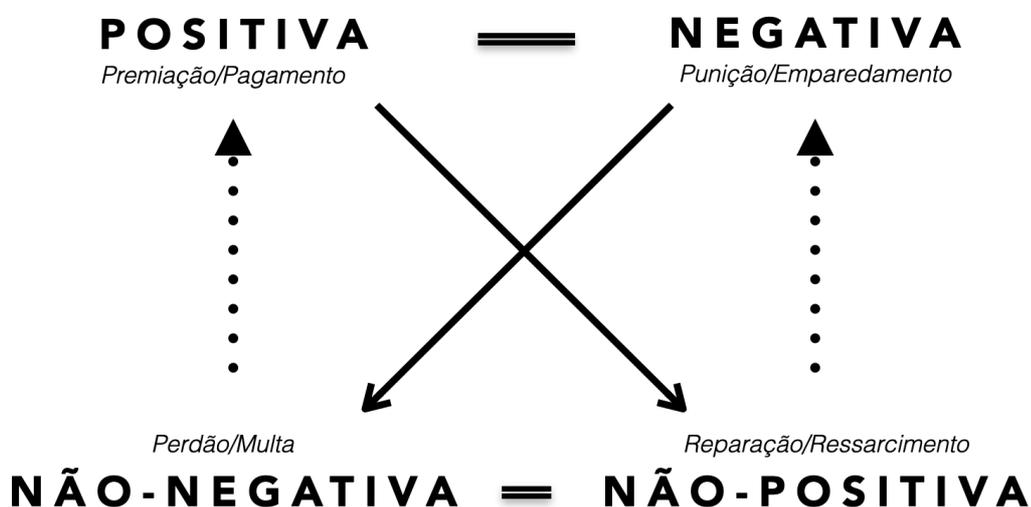
A sanção cognitiva centra-se no ser do sujeito; mais especificamente, sobre o estado decorrente de uma transformação. Ser e parecer, portanto, são os principais operadores deste tipo actancial. Se pensarmos em nosso objeto, uma avaliação de restaurante, as análises que levam em conta tal sanção partem da relação de junção entre o sujeito de estado e o objeto-refeição. Um sujeito que parece, mas que não está saciado, por exemplo, poderia configurar-se como um sujeito da conformação. Enquanto que um outro, que está saciado, mas não parece (pois não para de comer), poderia se configurar como um sujeito da compulsão. Acreditamos que a sanção cognitiva ajude a construir tipologias de sujeitos alimentados. Entretanto, no atual momento desta pesquisa, nos parece que a maior pertinência da sanção recai sobre o fazer do sujeito operador.

⁵⁷ Nele vemos a oposição principal, formada pelas categorias do /ser/ e do /parecer/. A partir delas, por uma operação de contradição, obtemos o /não ser/ e o /não parecer/. A partir da estrutura, podemos obter um quadrado de segunda geração em que: a categoria da /verdade/ aparece como o termo complexo resultante da soma de /ser/ e /parecer/; já o /segredo/ seria algo que é (/ser/), mas não parece (/não-parecer/); enquanto que a /falsidade/ nem parece (/não-ser/) e nem é (/não-ser/); e, por fim, a /mentira/ não é (/não-ser/), mas parece (/parecer/)

A sanção pragmática é um juízo epistêmico, proferido pelo Destinator-julgador sobre a conformidade dos comportamentos e, mais precisamente, do programa narrativo do sujeito "performante" em relação ao sistema axiológico (de justiça, de "boas maneiras", de estética, etc.), implícito ou explícito, pelo menos tal como foi atualizado no contrato inicial. Do ponto de vista do Destinatário-sujeito, a sanção pragmática corresponde à retribuição: enquanto resultado, esta é a contrapartida, na estrutura da troca, exigida pela performance que o sujeito realizou de acordo com suas obrigações contratuais; pode ser positiva (recompensa) ou negativa (punição); neste último caso, conforme seja a punição aplicada por um Destinator individual ou social, a retribuição negativa se chamará vingança ou justiça. (GREIMAS e COURTÉS; [1979] 2016: 426)

Parece claro que é sobre isso que estamos falando quando investigamos o objeto avaliação de restaurante e que, portanto, seu universo semântico pode ser definido pelo quadrado exposto na figura 9, abaixo.

Figura 9 - Quadrado da sanção pragmática: as setas cheias representam a relação de contraditoriedade; já as pontilhadas, a implicação; enquanto que as linhas duplas, a contrariedade.



Fonte: produção própria

O quadrado semiótico, inspirado no quadrado lógico de Aristóteles, foi proposto por Greimas, em parceria com François Rastier, no artigo "O jogo das restrições semióticas", de 1968⁵⁸.

Grosso modo, tal dispositivo tenta delinear todos os possíveis a partir de uma série de enquadramentos lógicos-semânticos. Seria como pensar que, fora dele, o homem não é capaz de organizar o sentido. Para construir um quadrado, em primeiro lugar se utilizam dois conceitos opostos, /negativo/ e /positivo/, por exemplo, que são os contrários. Então, na parte de baixo, por uma operação de contradição, se estabelecem os sub-contrários /não-positivo/ e /não-negativo/. Uma relação de implicação, por fim, é estabelecida entre o sub-contrário e o contrário oposto ao qual ele se origina: /não-positivo/ implica /negativo/. Inúmeras pesquisas dão conta de certo constrangimento causado por essa implicação. Entretanto, em *Tensão e Significação*, encontramos um ótimo resumo do problema.

Se a implicação constitui uma dificuldade, é aparentemente por supor uma homogeneidade da categoria que a contradição, por sua vez, coloca em xeque, negando o eixo semântico e abrindo uma infinidade de possíveis capazes de desestabilizar a categoria⁵⁹. Para que "não-pobre" possa implicar "rico", parece necessário, segundo essa objeção, postular-se antes de mais nada uma redução de todos os gradientes subjacentes, para se atingir o menor número possível de posições. A objeção de B. Pottier recai justamente, entre outras coisas, sobre a legitimidade de tal redução, na medida em que há uma infinidade de maneiras de não ser "pobre", a maioria das quais não consiste em ser "rico". (FONTANILLE e ZILBERBERG; [1998] 2001: 71-72)

Para resolver esse problema, o semioticista Claude Zilberberg apresenta o conceito de rede.

No estudo intitulado *Structure générale des corrélations linguistiques*, Hjelmslev considera que o paradigma da análise compreende duas possibilidades: a "análise por dimensões" e a "análise por subdivisão" (1985: 49)⁶⁰. A primeira, que conta

⁵⁸ GREIMAS, A.; RASTIER, F.; **The interaction of semiotic constraints**; Yale French Studies, n 41; EUA. Disponível em GREIMAS, A.; **Sobre o sentido: ensaios semióticos**; Editora Vozes; Petrópolis; [1970] 1975; pp 126-143. Trad. Ana Cristina Cruz Cezar (e outros).

⁵⁹ Nota de FONTANILLE e ZILBERBERG: GREIMAS, A.; FONTANILLE, J.; **Le beau geste**, RSSI (*Recherches Sémiotiques/ Semiotic Inquiry*), 13, 1-2, Canadá; 1993; pp 21-35.

⁶⁰ Citação de ZILBERBERG: HJELMSLEV, L.; **Nouveaux essais**; PUF; Paris; 1985.

visivelmente com a preferência do autor, “consistiria em reconhecer, no interior de uma categoria, duas ou mais subcategorias que se entrecruzam ou se interpenetram”. As resultantes são distintas: “Numa palavra: de acordo com a análise por dimensões, as subcategorias formam uma rede; de acordo com a análise por subdivisão, as subcategorias formam uma hierarquia”. Vejamos a mais simples das redes:

	c ↓	d ↓
a →	$A \approx ac$	$B \approx ad$
b →	$C \approx bc$	$D \approx bd$

(2011[2006]: 275)

Em nosso estudo, pretendemos desdobrar uma rede a partir das sanções /negativa/ e /positiva/ conforme proposto no *Dicionário de Semiótica* citado acima: ou seja, a partir dos termos /punição/ e /premiação⁶¹/, como contrários, e /perdão/ e /reparação/ como subcontrários. Para isso, além da categoria da positividade versus negatividade, também consideraremos a tonicidade dos termos. Obtendo, portanto, a rede da sanção (figura 10).

⁶¹ Optamos pelos termos premiação, reparação, punição e perdão para manter certa homogeneidade entre os conceitos. Por conta disso, deixamos recompensa, conforme proposta do *Dicionário de Semiótica*, de lado. Lembrando que a nomenclatura diz menos do que os fúntivos que a formam.

Figura 10 - Rede da sanção. Conforme bem salientado por Zilberbeg: “Não estamos considerando aqui os termos, e sim as relações constituintes. Ou melhor: é pelas definições que nos interessamos, e não pelas denominações que as subsumem.” (ZILBERBERG; [1988] 2006: 66)

Sanção Tonicidade	Positividade	Negatividade
Tônico	Premiação	Punição
Átono	Perdão	Reparação

Fonte: produção própria

Acreditamos, portanto, que o diagrama em rede seja preferível, na hora de evidenciar e de delinear o fenômeno, ao diagrama em quadratura, pois evidencia os gradientes subjacentes. Entretanto, a operacionalidade da rede para uma análise não parece ser muito eficiente. É como se pensássemos: “tudo bem, mas e o que eu faço com isso?” Unidos de um quadrado semiótico, por exemplo, inúmeras análises dão conta de explicar fenômenos textuais em que se passa, por exemplo, do tema da vida para o da morte⁶². Como passar, então, do pagamento (premiação) para o ressarcimento⁶³ (reparação), que é o não-pagamento, o “quero meu dinheiro de volta”, constituindo uma sintaxe?⁶⁴

Zilberberg utiliza um plano cartesiano em que imediatamente dispõe todos os termos. Entretanto, acreditamos que é o modelo catenário de Bevidas (figura 11) a melhor solução para fazer ver todo o universo do sentido que a sanção pode comportar. Nele, parece ficar claro que não se passa da premiação para a punição, mas sim para a reparação. Da mesma forma que não se passa da punição para a premiação, mas sim para o perdão. São universos com forte afinidade entre si,

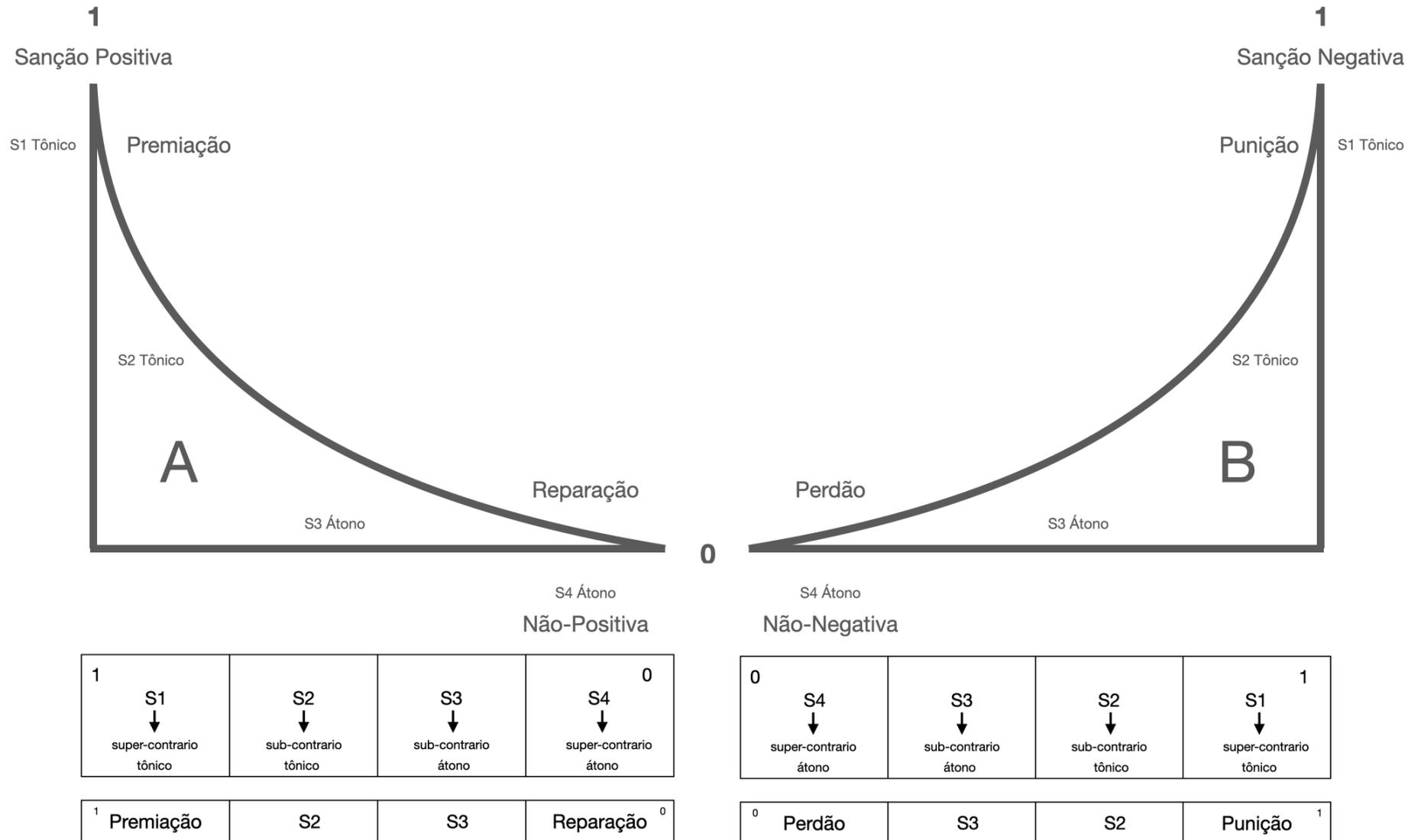
⁶² Mencionado em 1.0 a respeito do trabalho de Arnold junto a análise de *A máquina mole*.

⁶³ Pagamento, ressarcimento, emparedamento, multa, indulto etc., são figurativizações de nossos termos primários: premiação, reparação, punição e perdão.

⁶⁴ A pertinência de tais questões se esclarecerá abaixo, na apresentação dos textos relativos a cada sanção, e também no modelo empregado para a classificação de nossos textos, no capítulo três.

Figura 11 - Modelo catenário e tensivo da sanção pragmática

66



Fonte: criação própria a partir de Bevidas (2019; e [no prelo])

entretanto, a rede nos faz ver que a tonicidade marca uma forte diferenciação.

Nossa hipótese, portanto, e, aparentemente em bastante afinidade com o que postula o *Dicionário de Semiótica*, é a de que o TripAdvisor é sempre sanção positiva ou não-positiva. Discursos de sanção negativa, a nosso ver, estão sempre relacionados à esfera jurídica. Das leis institucionalizadas, no caso da justiça, ou da ruptura com esses contratos, no da vingança.

Por exemplo, o projeto de lei 01-00099/2019, do vereador paulistano Reginaldo Tripoli (PV), que versa sobre a proibição de fornecimento de produtos de plástico em bares restaurantes, apresenta as sanções que um restaurante poderá receber ao infringir o contrato legal. No artigo 5º ficam estabelecidos seis incisos em que as punições são estabelecidas gradualmente. Como não poderia deixar de ser, a punição mais tônica é a última a aparecer.

VI - se desrespeitado o fechamento administrativo, será requerida a instauração de inquérito policial, com base no art. 330 do código Penal, e realizado novo fechamento ou embargo de obra, com auxílio policial, se necessário, e, a critério da fiscalização, poderão ser utilizados meios físicos que criem obstáculos ao acesso, tais como emparedamento, defensas de concreto, tubos de concreto, dentre outros.⁶⁵

Quando falamos, em pleno Século XXI, sobre a existência de punição por emparedamento, não são poucas as pessoas que expressam consternação. Entretanto, o impedimento de que um restaurante volte a estabelecer relações de conjunção entre um cliente e uma refeição é definido em lei. Mais do que isso, o discurso legal ainda estabelece gradações, ou seja, nos incisos anteriores ao que estabelece o emparedamento, é possível encontrar autuações, advertências e multas. Outros tipos de leis, entretanto, falam sobre o perdão. É o caso do parágrafo XII do artigo 84 da Constituição Federal⁶⁶, que estabelece a competência do presidente da república de estabelecer indultos e comutações de penas.

⁶⁵ Disponível em: <http://documentacao.saopaulo.sp.leg.br/iah/fulltext/projeto/PL0099-2019.pdf>
Consulta: 19/01/2021.

⁶⁶ BRASIL [Constituição (1988)]; **Constituição da República Federativa do Brasil: texto constitucional promulgado em 5 de outubro de 1988, com as alterações adotadas pelas Emendas constitucionais n 1/1992 a 92/2016, pelo Decreto legislativo n 186/2008 e pelas Emendas constitucionais de revisão n 1 a 6/1994**; Brasília; Edições Câmara; 2016.

Tudo isso dito apenas para demonstrar que o universo do sentido é amplo e precisa ser suficiente e pertinentemente bem delineado para evitar confusões. Ou seja, como dito anteriormente, sanção negativa não parece fazer parte do nosso corpus, uma vez que diz respeito apenas ao universo jurídico - inclusive no que tange ao desrespeito a esse⁶⁷.

Entretanto, há uma exceção que não funciona exatamente desta forma: o boicote. Em se tratando dos netizens (principalmente), tal prática, em sua forma discursiva, vem sendo caracterizada como "tribunal da internet" ou "cultura do cancelamento". Na figura 12, logo abaixo, temos o exemplo de um boicote encontrado em nosso objeto de estudo. No texto, proferido por um usuário do TripAdvisor, o empresário Junior Durski, dono de uma rede de hamburguerias, é chamado de genocida. Mais do que isso, o ator da enunciação ainda exige que outros usuários eliminem "o restaurante do nosso mapa". Apesar de nosso modelo de classificação dar conta do fenômeno, pois estabelece a categoria da extinção, ele não é exatamente o mesmo tipo de sanção que estamos trabalhando.

Figura 12 - Avaliação de restaurante com caráter de boicote⁶⁸.



crismadf
2 avaliações

●○○○○○ Publicada em 24 de março de 2020

Dono genocida

Temos o dever cívico e humanitário de eliminar este restaurante de nosso mapa. Dono sem noção. Declarações que mostram qual o valor e missão da empresa. Está é minha opinião.

Data da visita: março de 2020

Útil?

Fonte: reprodução/ TripAdvisor

Para estabelecer a diferença entre programas de uso e programas de base, o *Dicionário de Semiótica* (GREIMAS e COURTÉS; [1979] 2016: 527) vale-se do

⁶⁷ Em novembro de 2020, manifestantes durante protesto contra a morte de um homem em loja da rede Carrefour de supermercados atearam fogo no estabelecimento. Confira em: <https://www.metropoles.com/brasil/em-protesto-contr-o-racismo-manifestantes-tocam-fogo-em-carrefour-de-sp>. Consulta: 01/02/2021.

⁶⁸ A tentativa de boicote, realizada pelo ator da enunciação, parece referir-se ao fato de que Junior Durski, dono da rede de hamburguerias Madero, minimizou o impacto da pandemia de Covid-19 com discurso em prol da economia: "Brasil não pode parar por 5 ou 7 mil mortes". Confira em: <https://congressoemfoco.uol.com.br/economia/brasil-nao-pode-parar-por-cinco-ou-sete-mil-mortes-diz-dono-do-madero/>. Consulta: 01/02/2021.

famoso exemplo do macaco e da banana. Greimas pontua tal diferença como forma de demonstrar a capacidade de a semiótica discursiva ocupar-se de narrativas complexas. Sendo assim, quando vemos um macaco com um pedaço de pau na mão, tentando derrubar um cacho de banana, o jogo das pressuposições é rapidamente acionado. Para o mamífero, o toco de madeira representa um valor de uso, enquanto que a banana é o valor de base. Partindo dessa analogia, é fácil observar que para o usuário que solicita um boicote, a sanção é um programa de uso que visa o programa de base "eliminar este restaurante de nosso mapa". Ele não está ali para avaliar a comida, o serviço e nem nada. A transformação na relação de junção entre sujeito de estado e objeto refeição pouco lhe importa. Sendo assim, há uma extinção do programa que é base no TripAdvisor em prol de outro, que só pertence ao usuário (ou aos demais que ele conseguir amearhar nesse caminho).

Antes de chegarmos à análise do TripAdvisor, propriamente dito (coisa que faremos a partir de 3.2), cabem aqui mais algumas linhas sobre o Reclame Aqui. Plataforma que, ao nosso ver, é do mesmo tipo que aquelas de avaliações de restaurantes. Entretanto, coerções de vários tipos (inclusive da própria arquitetura de dados das empresas digitais) parecem centralizar os discursos do pagamento no TripAdvisor e os discursos do ressarcimento no Reclame Aqui. É difícil generalizar esse tipo de coisas, principalmente em se tratando de semiótica, mas a forma como as plataformas orientam-se discursivamente - o Reclame Aqui se colocando no papel de site que auxilia a obtenção do ressarcimento, enquanto que no TripAdvisor esse lado advocatício nem se menciona -, parecem direcionar os usuários a produzir mais textos de pagamento em uma e mais textos do ressarcimento em outra. É inegável que existem textos do tipo "quero meu dinheiro de volta" na plataforma do TripAdvisor, coisa que se verifica, inclusive, em nosso corpus. Mas a quantidade dos textos de pagamento é fortemente mais presente.⁶⁹

⁶⁹ Tal fato será bastante evidenciado no próximo capítulo e no apêndice, inclusive a respeito de um sobreajuste (overfitting) causado em nosso modelo de inteligência artificial devido ao desbalanceamento do corpus: mais de 80% das avaliações classificadas manualmente eram discursos do pagamento.

3. MANIFESTAÇÃO: NOVOS ACERCAMENTOS À SUPERFÍCIE DO TEXTO

Neste capítulo, iremos nos centrar na superfície do texto. Saindo de níveis abstratos, como fizemos até agora, finalmente iremos tocar nas palavras dispostas pelos usuários do TripAdvisor. Como o título indica, pretendemos tratar de acercamentos relativamente novos, principalmente em se tratando da área da semiótica e da análise do discurso.

Na primeira seção, apresentaremos brevemente o conceito de análise de sentimento: os principais métodos computacionais de implementação da técnica e algumas dificuldades já enfrentadas pelos utilizadores dela. Aqui, vale ressaltar, faremos uma avaliação puramente descritiva com base em artigos e livros, não iremos utilizar a semiótica para analisar o modelo.

Depois, em 3.2, apresentaremos a proposta das “quantificações subjetivas”, criação de Luiz Tatit com base na hipótese tensiva de Claude Zilberberg. Nessa seção, debateremos brevemente o conceito e alguns de seus desdobramentos recentes. Imediatamente a seguir, utilizaremos uma sub-seção para demonstrar a aplicabilidade do conceito. Para isso, usaremos cada um de seus componentes na classificação de nosso corpus.

Por fim, na última seção (3.3) apresentaremos resultados de nossa prova de conceito. Para isso, discorreremos brevemente sobre o conceito de neurônio e de redes neurais.

3.1. Análise de sentimento

Para introduzir uma das principais obras de referência no processamento de linguagem natural, Daniel Jurafsky e James Martin (2000) utilizam o personagem HAL 9000, uma inteligência artificial presente no filme “2001: Uma Odisseia no Espaço”, de Stanley Kubrick. Por meio da obra, os pesquisadores definem a capacidade de um autômato interagir com seres humanos por meio de inúmeros componentes computacionais: reconhecimento de fala, entendimento de linguagem

natural, leitura labial, geração de linguagem natural, síntese de fala, recuperação de informação, inferência e extração de informação. Segundo os pesquisadores, Arthur C. Clarke, quando criou o personagem HAL, estava muito otimista em relação a quanto tempo um autômato como aquele estaria disponível. Afinal, ainda hoje (20 anos após a data prevista), muitos dos problemas por trás de cada um daqueles componentes (necessários para que uma inteligência artificial se comunique com humanos da mesma forma como apresentado na película) ainda estão bastante longe de serem resolvidos computacionalmente.

Para dar conta dessas soluções, entretanto, as ciências da computação não são capazes de operar sozinhas. Para JURAFSKY e MARTIN (2000), isso é responsável por configurar o campo multidisciplinar de áreas como Processamento de Linguagem Natural, Linguística Computacional e Reconhecimento e Síntese da Fala (que poderiam ser agrupados em uma grande disciplina conhecida como Processamento de Língua e Fala). Todas elas são empregadas nas mais diversas atividades desde "contagem de palavras e hifenização automática, a aplicativos de ponta como os de produção automática de respostas na internet e tradutores de linguagem falada em tempo real." (JURAFSKY e MARTIN; 2000: 2)⁷⁰

Mais próximos do tema de nossa pesquisa, os pesquisadores iniciam a terceira parte do livro, dedicada à semântica, explicando aquilo que Madeleine Arnold chama de representação do conhecimento. Para JURAFSKY e MARTIN (2000), entretanto, o conceito se torna "representação da significação". Epistemologicamente muito distantes de nossos fundamentos (uma vez que buscamos nos manter o tempo inteiro em imanência), os pesquisadores apresentam algumas atividades computacionais que, segundo eles, envolvem um "conhecimento não linguístico do mundo"⁷¹ (JURAFSKY e MARTIN; 2000: 498), ou seja, um conhecimento transcendente à linguagem.

Ler um cardápio e decidir o que pedir, dar conselhos sobre onde jantar, seguir uma receita e gerar novas receitas exigem um conhecimento profundo sobre os

⁷⁰ Tradução nossa para: "word counting and automatic hyphenation, to cutting edge applications such as automated question answering on the Web, and real-time spoken language translation."

⁷¹ Tradução nossa para: "non-linguistic knowledge of the world"

alimentos, o modo de preparo deles, o que as pessoas gostam de comer e como são os restaurantes.⁷² (JURAFSKY e MARTIN; 2000: 498)

Entretanto, basta ler a análise da sopa ao pesto feita por Greimas, já mencionada algumas vezes neste trabalho, para perceber que a semiótica propõe maneiras não transcendentais para se debruçar sobre tais temas. Longe de criticar a pesquisa fundamental para o processamento de linguagem natural, fornecida por JURAFSKY e MARTIN (e utilizada por praticamente todos os especialistas em PLN do planeta), essas primeiras linhas apresentam um distanciamento fundamental entre nossa disciplina e a linguística dos pesquisadores americanos.

Na abordagem representacional que está sendo explorada aqui, pegamos entradas linguísticas e construímos representações de significado que são feitas do mesmo tipo de material que é usado para representar esse tipo de conhecimento do mundo cotidiano pelo senso comum. O processo pelo qual tais representações são criadas e atribuídas a entradas linguísticas é chamado de análise semântica.⁷³ (JURAFSKY e MARTIN; 2000: 498)

Epistemologicamente, nós temos um problema bastante sério. Afinal, o senso comum é sempre muito falho na construção científica, fato largamente explorado por Gaston Bachelard em *A formação do espírito científico*. Apenas para ficarmos com a epítome que inicia esta dissertação, “A dualidade Universo e Espírito, quando examinada no âmbito de um esforço de conhecimento pessoal, aparece como a dualidade do fenômeno mal preparado e da sensação não retificada.” (BACHELARD; 1996: 297)

Utilizar o senso comum para construir representações do Universo sempre esbarra em questões de subjetividade e de inferências mal realizadas. Não faltam exemplos na história das ciências: determinismo geográfico, positivismo evolucionista de Lombroso, entre outros, são testemunhos das escorregadelas cognitivas de quando a sensação não se retifica e o fenômeno não é bem preparado para análise.

⁷² Tradução nossa para: “Reading a menu and deciding what to order, giving advice about where to go to dinner, following a recipe, and generating new recipes all require deep knowledge about food, its preparation, what people like to eat and what restaurants are like.”

⁷³ Tradução nossa para: “In the representational approach being explored here, we take linguistic inputs and construct meaning representations that are made up of the same kind of stuff that is used to represent this kind of everyday common-sense knowledge of the world. The process whereby such representations are created and assigned to linguistic inputs is called semantic analysis.”

Entretanto, apesar da abertura para críticas que as citações anteriores se dispõem, elas servem a um propósito mais aquém à nossa pesquisa: lembrar que Madeleine Arnold postulava que duas contribuições da semiótica já haviam se demonstrado bastante frutíferas junto a atividades computacionais. Uma delas, e a mais importante para nós, é a de trabalhar junto a “sistemas cujos modos de raciocínio e expressão são acessíveis ao usuário e o mais próximos possível aos dos seres humanos” (ARNOLD; 1989: 86).

De acordo com JURAFSKY e MARTIN (2000), e inspirados pelo “poder otimizador” da semiótica, acreditamos poder afirmar que todos os trabalhos em processamento de linguagem natural que se debruçam sobre análise semântica⁷⁴ estão abertos para novos estudos de um ponto de vista imanente à linguagem⁷⁵. Dentro daquele campo, encontra-se a análise de sentimento.

Um número considerável de artigos que mencionam “análise de sentimento” concentra-se na aplicação específica de classificações de avaliações quanto à sua polaridade (positiva ou negativa), um fato que parece ter levado alguns autores a sugerir que a frase se refere especificamente a esta tarefa estritamente definida. No entanto, hoje em dia, muitos interpretam o termo de forma mais ampla para significar o tratamento computacional de opinião, sentimento e subjetividade no texto. (PANG e LEE; 2008: 12)⁷⁶

O campo do processamento computacional da subjetividade é bastante novo e ainda não formalizado. Conforme o excerto acima, existe um guarda-chuva maior, debaixo do qual foram desenvolvidas inúmeras metodologias para se abordar o texto. Atualmente, até a análise do ‘tom’ de um enunciado (irônico, engraçado etc.) tem sido proposto em inúmeros artigos. Entretanto, devido ao escopo deste trabalho, nos debruçaremos apenas sobre as atividades computacionais que classificam avaliações quanto à polaridade (negativa ou positiva).

⁷⁴ Confira as partes III e IV de JURAFSKY e MARTIN (2000: 495-822) para mais informações sobre o fértil campo do processamento de linguagem natural que está aberto à semiótica.

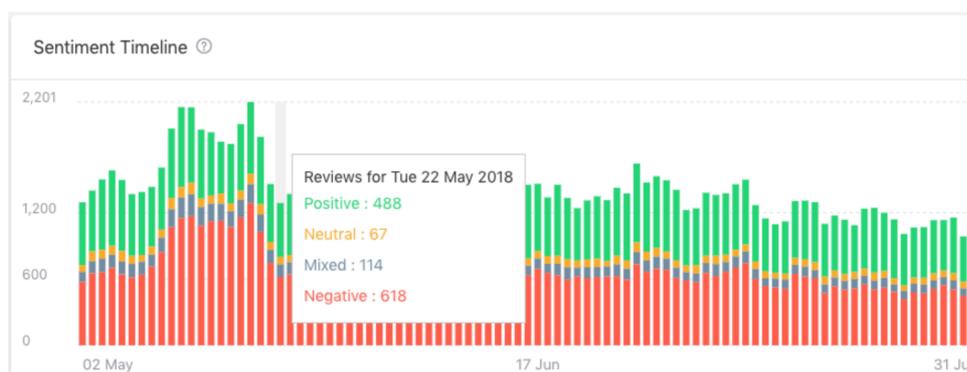
⁷⁵ Para uma versão resumida da relação entre semiótica e imanência da linguagem, confira BEIVIDAS, W.; **A semiótica tensiva: uma teoria imanente ao afeto**; 2015, p. 43-86

⁷⁶ Tradução nossa para: “A sizeable number of papers mentioning “sentiment analysis” focus on the specific application of classifying reviews as to their polarity (either positive or negative), a fact that appears to have caused some authors to suggest that the phrase refers specifically to this narrowly defined task. However, nowadays many construe the term more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text”

Uma tecnologia fundamental em muitas aplicações atuais de mineração de opinião e análise de sentimento é a classificação(...) A razão pela qual a classificação é tão importante é que muitos problemas de interesse podem ser formulados a partir da aplicação de classificação/regressão/ranqueamento a determinadas unidades textuais; exemplos incluem decisões a partir de uma frase ou documento específico (“quão positivo é?”), ordenar um conjunto de textos (“classifique essas avaliações pela sua positividade”)... (PANG e LEE; 2008: 23)⁷⁷

Na figura 13 (abaixo), encontramos um exemplo de representação gráfica da análise de sentimento fornecida pela plataforma Appbot, que oferece monitoramento e relatórios para empresas. Nela, é possível ver quatro classes de classificação: positiva, neutra, negativa e misturada. “O Appbot usa um algoritmo patenteado com tecnologia de IA para determinar o sentimento de cada avaliação. A análise de sentimento do Appbot faz referência principalmente ao conteúdo do texto da avaliação.”⁷⁸ Ainda segundo a empresa, a notação faz uso do senso comum para afirmar que as classes positiva e negativa referem-se a avaliações em que uma dessas categorias esteve mais presente. Quanto às neutras, não haveria um posicionamento claro neste sentido e, em relação à misturada, ambas as classes primárias estariam disponíveis nas mesmas proporções.

Figura 13 - Tela de apresentação de análise de sentimentos do Appbot



Fonte: reprodução/ Appbot

⁷⁷ Tradução nossa para: “A fundamental technology in many current opinion-mining and sentiment-analysis applications is classification(...) The reason that classification is so important is that many problems of interest can be formulated as applying classification/regression/ranking to given textual units; examples include making a decision for a particular phrase or document (‘how positive is it?’), ordering a set of texts (‘rank these reviews by how positive they are’)”

⁷⁸ Tradução nossa para: “Appbot uses a proprietary AI-powered algorithm to determine the sentiment of each review. Appbot’s sentiment analysis primarily references the text content of the review.” Disponível em: <https://support.appbot.co/help-docs/positive-sentiment-negative-sentiment/>. Consulta: 20/01/2021.

A partir de um corpus classificado com sentenças positivas e negativas⁷⁹, as metodologias empregadas fazem uso de sumarização. Ou seja, resumem um texto ou segmento de texto àquelas categorias. Para isso, lógicas diferentes de algoritmos computacionais estão disponíveis. Segundo revisão de DONEGA (2017: 50), as principais metodologias utilizadas para realizar tal tarefa de classificação utilizam aprendizado de máquina supervisionado ou dicionários e ontologias. Neste segundo caso, que não faz parte do escopo deste trabalho, são criadas regras pelo programador por meio de combinações de palavras que podem ou não estar presentes e de como se comportam em presença de umas e de outras.

Em aprendizado de máquina, um dos principais mecanismos de inteligência artificial é conhecido como classificador bayesiano ingênuo. Segundo Ferreira e Lopes (2019: 164-169), é um modelo de classificação que, a partir da regra de Bayes, calcula probabilidades condicionadas de ocorrências.

(...)o modelo subjacente a um classificador bayesiano ingênuo parte do princípio de que os atributos são condicionalmente independentes. Em outras palavras, a ocorrência de um atributo f_i em um dado pertencente a uma certa classe não afeta a probabilidade de um outro atributo f_j ocorrer nesse mesmo dado. (FERREIRA e LOPES; 2019: 169)

Para exemplificar o conceito, os pesquisadores utilizam um classificador de e-mails entre spam e não spam (categoria binária que faz as vezes de positivo e negativo neste caso). A partir apenas do corpo do texto, sem considerar nenhum metatexto, eles propõem utilizar a ocorrência, ou não, de certas palavras para determinar a classe a que o texto faz parte.

(...)vamos considerar o vocabulário como uma lista indexada de palavras pré-selecionadas ($palavra_1$, $palavra_2$, ..., $palavra_n$). Dessa forma, toda mensagem será representada por n atributos, cada um deles (f_i) indicando a ocorrência ou não de $palavra_i$ na mensagem. (FERREIRA e LOPES; 2019: 170)

Com fins didáticos, para demonstrar o funcionamento do classificador bayesiano ingênuo, Ferreira e Lopes restringem os atributos a apenas 3 palavras: *viagra*, *sexo* e *reunião*. Depois disso, simulam um corpus de treinamento com 10 mil mensagens

⁷⁹ A atividade pode ser feita manualmente ou, em alguns casos, há a tentativa de distribuir notação simbólica (como as estrelas para restaurante, em que 1 significa péssimo e 5, excelente) nas classes positiva e negativa.

já classificadas, em que seis mil são spam e quatro mil não spam. Então, apresentam tabelas com os valores relativos à presença das palavras selecionadas em cada uma das categorias. Por fim, apresentam previsões de novos textos pertencerem a uma classe ou outra por meio da combinação da existência ou não de cada uma das palavras no texto em questão.

Nesse momento, talvez você desconfie da hipótese de independência condicional entre os atributos usados na classificação. Afinal, não parece muito realista, por exemplo, pensar que a ocorrência da palavra “viagra” numa mensagem de spam não afeta a probabilidade da palavra “sexo” ocorrer na mesma mensagem, quando sabemos que essas duas palavras estão muito frequentemente associadas. Ainda assim, a simplificação computacional resultante é valiosa e pode compensar o fato de essa associação entre as palavras ser deixada de fora. (FERREIRA e LOPES; 2019: 174)

Apesar dessa ingenuidade, de utilizar probabilidade condicionada para representar o universo, o método bayesiano, ainda hoje, é uma das principais ferramentas de aprendizado de máquina. Entre os motivos para isso estão a fácil implementação de códigos e o “bom custo-benefício, aliando simplicidade a bons resultados” (FERREIRA e LOPES; 2019: 183). Apesar disso, optamos por utilizar *deep learning* na prova de conceito de nosso classificador também devido à facilidade de implementação do código (principalmente a partir da biblioteca *Fast.AI*), mas principalmente pelo interesse de iniciar nossos estudos nas redes neurais⁸⁰. Para uma validação mais completa de nossa proposta de classificação, sabemos que será necessário comparar com os diversos outros métodos de PLN. Por ora, vamos apresentar algumas características de *deep learning*.

As abordagens usadas para análise de sentimentos são numerosas e são baseadas em diferentes métodos de processamento de linguagem natural e técnicas de

⁸⁰ Conforme já mencionamos ao longo da dissertação - e evidenciaremos melhor a seguir e em 3.3 -, as estratégias e os algoritmos de *deep learning* parecem já dar conta de termos anafóricos e catafóricos, além de responderem ao problema, levantado por Arnold, sobre o reconhecimento de formas por meio das inteligências artificiais. Nossa hipótese, portanto, é a de que as normalizações realizadas no corpus para produzir maior ordem, como lematização e limitação do vocabulário, causam perdas semânticas que podem afetar no resultado de classificação quando um corpus é composto por textos como os das redes sociais digitais. O método bayesiano, ao que tudo indica, tem ótima funcionalidade apenas em textos próximos à norma culta. Destarte, futuramente, poderemos comparar resultados e checar a validade da hipótese. Sobre os problemas entre norma culta e linguagem na internet, vale conferir a proposta do português-twitter (ASTI, P.; **Anotador morfossintático para o português-twitter**; Rio de Janeiro: PUC-Rio, Departamento de Informática, 2011.) e a da complexidade discursiva (BARROS, D.; **A complexidade discursiva na internet**, CASA: Cadernos de Semiótica Aplicada, v.13, n.2, 2015, p. 13-31)

aprendizado de máquina para extrair recursos adequados e classificar texto em rótulos de polaridade apropriados. Desde alguns anos, com a popularidade que os métodos de *deep learning* ganharam, várias redes neurais profundas foram utilizadas no campo com sucesso. Particularmente, as redes neurais convolucionais e as redes LSTM provaram ter um bom desempenho para tarefas de análise de sentimento.⁸¹ (GOULARAS e KAMIS; 2019: 12)

As redes neurais são consideradas a vanguarda em tarefas de inteligência artificial nos dias de hoje⁸². Mais do que o valor já dado pelas ciências da computação, acreditamos que a utilização de redes neurais possa resolver o problema da elasticidade da língua e das regras que precisariam ser criadas uma a uma pelo analista. Ao apresentarmos dados rotulados para o computador, ele aprende por conta própria.

Essa solução [para resolver os problemas mais difíceis da inteligência artificial, como reconhecimento de palavras e de rosto em uma imagem] permite que os computadores aprendam com a experiência e entendam o mundo em termos de uma hierarquia de conceitos, com cada conceito definido em termos de sua relação com conceitos mais simples. Ao reunir conhecimento da experiência, essa abordagem evita a necessidade de operadores humanos especificarem formalmente todo o conhecimento que o computador precisa. A hierarquia de conceitos permite que o computador aprenda conceitos complicados construindo-os a partir de outros mais simples. Se desenharmos um gráfico mostrando como esses conceitos são construídos uns sobre os outros, o gráfico será profundo, com muitas camadas. Por esse motivo, chamamos essa abordagem de IA de *deep learning*.⁸³ (GOODFELLOW e BENGIO; 2016: 1-2)

⁸¹ Tradução nossa para: "The approaches used for sentiments analysis are numerous and are based on different methods of natural language processing and machine learning techniques for extracting adequate features and classifying text in appropriate polarity labels. Since some years, with the popularity that deep learning methods have gained, various deep neural networks were utilized on the field with success. Particularly, the convolutional neural networks and LSTM networks proved to be performant for sentiment analysis tasks."

⁸² Em 3.3 apresentamos o conceito de neurônio e utilizamos um modelo de rede neural para realizar nossa prova de conceito.

⁸³ Tradução nossa para: "This solution is to allow computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined in terms of its relation to simpler concepts. By gathering knowledge from experience, this approach avoids the need for human operators to formally specify all of the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason, we call this approach to AI deep learning."

Madeleine Arnold já havia notado a dificuldade de os computadores reconhecerem formas automaticamente, conforme apresentamos em 1.3. Naquela época, entretanto, as inteligências artificiais eram tão incipientes, que sequer haviam obtido sucesso no xadrez⁸⁴.

Ironicamente, tarefas abstratas e formais, que estão entre as atividades mentais mais difíceis para um ser humano, estão entre as mais fáceis para um computador. Os computadores há muito tempo são capazes de derrotar até mesmo o melhor jogador de xadrez humano, mas só recentemente estão igualando algumas das habilidades comuns dos seres humanos de reconhecer objetos ou fala.⁸⁵ (GOODFELLOW e BENGIO; 2016: 2)

Tais dificuldades têm sido resolvidas com grande sucesso pelos algoritmos de *deep learning*. Entretanto, a tecnologia ainda bastante nova exige muito estudo para que um analista aprenda a operar seus hiperparâmetros.

De acordo com os autores deste livro, o aprendizado de máquina é a única abordagem viável para construir sistemas de IA que podem operar em ambientes complicados do mundo real. *Deep learning* é um tipo específico de aprendizado de máquina que atinge grande poder e flexibilidade ao aprender a representar o mundo como uma hierarquia aninhada de conceitos, com cada conceito definido em relação a conceitos mais simples e representações mais abstratas computadas em termos de outras menos abstratas.⁸⁶ (GOODFELLOW e BENGIO; 2016: 8)

Nossa intenção não é a de ser exaustivos, mas introduzir algumas questões bastante pertinentes e que já se encontravam em jogo desde os anos 1980. Entretanto, o conceito de *deep learning* (DL) e até mesmo de *machine learning* (ML), ainda não estavam presentes nos trabalhos de Arnold.

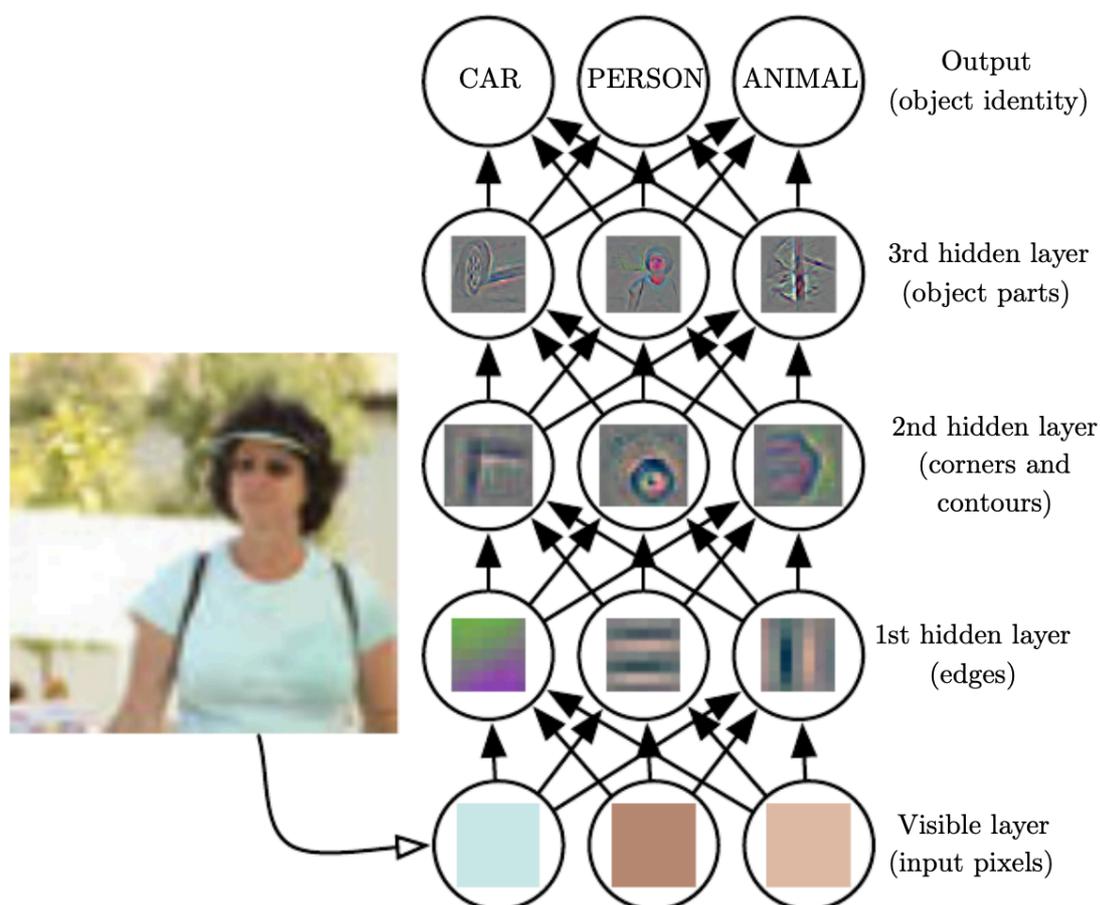
⁸⁴ O último artigo da semiótica, a respeito da semiótica como representação do conhecimento, data de 1989, enquanto que Garry Kasparov, o famoso enxadrista russo, só foi derrotado pela inteligência Deep Blue, da IBM, em 1997.

⁸⁵ Tradução nossa para: "Ironically, abstract and formal tasks that are among the most difficult mental undertakings for a human being are among the easiest for a computer. Computers have long been able to defeat even the best human chess player, but are only recently matching some of the abilities of average human beings to recognize objects or speech."

⁸⁶ Tradução nossa para: "According to the authors of this book, machine learning is the only viable approach to building AI systems that can operate in complicated, real-world environments. Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones."

No exemplo do detector de spam, temos um clássico método de ML, que segundo (PONTI e COSTA; 2017), é considerado como raso ou superficial, de acordo com a literatura. Para os autores, a principal diferença entre ML e DL resulta no fato de que no primeiro há a busca de uma função única que seja capaz, a partir de certos parâmetros (como a presença das palavras *viagra*, *sexo* e *reunião*), dar conta de representar a classe proposta. Agora, em relação a DL, as coisas mudam de grau, uma vez que tal mecanismo busca encontrar uma função que seja composta não de parâmetros, mas de outras funções de parâmetro.

Figura 14 - Multilayers ou função de funções



Fonte: reprodução (GOODFELLOW e BENGIO; 2016: 6)

Para entender melhor o funcionamento desse tipo de função de funções, GOODFELLOW e BENGIO fornecem a figura 14. Nela temos uma ilustração de como a máquina seleciona características importantes para diferenciar uma pessoa de um carro ou animal por meio das detecção de arestas, cantos e contornos e partes de objetos.

Para sermos mais precisos, cada função faz uso de parâmetros para realizar a transformação dos dados de entrada. Iremos denotar o conjunto desses parâmetros (comumente uma matriz) por W_l , relacionado a cada função f_l , e então podemos escrever:

$$f_L(\dots f_2(f_1(x_1, W_1); W_2) \dots), W_L),$$

onde x_1 representa os dados de entrada, cada função tem seu próprio conjunto de parâmetros e sua saída será passada para a próxima função. Na equação acima, temos a composição de L funções, ou L camadas. (PONTI e COSTA; 2017: 66)

Para encerrar e resumir a seção, cabe afirmar que a análise de sentimento nos ajudou a introduzir, brevemente, um problema que a teoria da linguagem de Hjelmslev considera como o inconveniente da indução. Nos *Prolegômenos a uma teoria da linguagem*, o mestre dinamarquês denuncia o “senso comum cientificado” ou método indutivo.

A experiência põe em evidência os inconvenientes de um tal método. Este conduz inevitavelmente à extração de conceitos hipostasiados como sendo reais. Esse realismo (no sentido medieval do termo) não fornece uma base utilizável para a comparação, dado que os conceitos assim obtidos não têm valor geral e só se aplicam a um determinado estágio de uma dada língua. A terminologia tradicional completa mostra o fracasso desse realismo: as classificações da gramática indutiva, tais como “genitivo”, “perfeito”, “subjuntivo”, “passivo” etc., são exemplos notáveis desse fato. (HJELMSLEV; [1943] 2018: 14)

Ou seja, apesar dos avanços tecnológicos e econômicos que incidem sobre a área, parece ser o senso comum que continua a ditar as regras para a criação de classes para as classificações. Sendo assim, é preciso lembrar que, já na década de 1960, Pierre Maranda alertava sobre a incapacidade de um computador fazer milagres a partir do *input* de “dados ruins”. Em se tratando de redes neurais, por outro lado, até existem exceções milagrosas, uma vez que a função de funções pode detectar que certas características originam-se de erro humano (*outliers*). Nossa hipótese, entretanto, ainda a ser verificada, é a de que, com uma notação mais acurada para a classificação, os sistemas performariam melhor⁸⁷.

⁸⁷ Como reforço à nossa hipótese, confira o artigo (já citado) de GOULARAS e KAMIS (2019). No trabalho, os autores denunciam a baixa performance da utilização de técnicas de *deep learning* para o processamento de linguagem natural em redes sociais digitais quando comparado a atividades de reconhecimento de imagens, entre outras tarefas computacionais.

Antes de iniciarmos a próxima seção, vale lembrar nossa demonstração a respeito da tipologia das sanções. Lá, deixamos claro que, semioticamente falando (e de um ponto de vista da ação), sanção negativa é sempre parte do universo jurídico. Tanto o TripAdvisor quanto o Reclame Aqui operam em uma direção tensiva que tem / Positivo/ (tônico) e /Não-Positivo/ (átono) como eixo norteador. Em 3.2. apresentaremos o modelo das quantificações subjetivas a ser acoplado nesta direção como proposta de otimização à análise de sentimento.

3.2. Quantificações subjetivas

O termo "quantificações subjetivas", em semiótica discursiva, foi uma criação brasileira. No trabalho da pesquisadora Paula Martins de Souza (2016: 57), afirma-se que a "expressão 'quantificações subjetivas' foi, de modo muito feliz, cunhada por Tatit" no artigo "Quantificações subjetivas: crônicas e críticas"⁸⁸

Segundo Souza, trata-se de um agrupamento de medidas que, tendo por base os incrementos zilberberguianos de "mais" e de "menos" permite aos semioticistas "depreender a significação da faceta subjetiva ao lado da objetiva" (SOUZA; 2016: 55). A pesquisadora exemplifica a utilização do modelo junto a significações passionais afirmando que, apesar de ser impossível metrificar numericamente a quantidade de amor, é possível, em uma comparação, verificar se a quantidade do sentimento é maior ou menor. Para Souza, Zilberberg utiliza as partículas "mais" e "menos" (também chamadas de "sílabas' tensivas" pelo francês ([2006] 2011: 253) como forma de metrificar uma gradação possível de se realizar em ambas as dimensões (objetiva e subjetiva) - metragem que concede a suas unidades uma direção e um posição aspectual.

Uma vez estabelecidos os limites das gradações, expressos por "só mais" e "só menos" na teoria, os limiares passam a ter valor posicional e direcional. As direções possíveis são duas: ou bem há um acréscimo crescente, constituindo uma ascendência, ou bem há um decréscimo, formando uma direção descendente. No caso da ascendência, as posições previstas, progressivamente, são o "só menos", o "menos menos", o "mais mais", culminando no "só mais". Na direção oposta, temos a progressão seguinte: "só mais", "menos mais", "mais menos" e "só menos". Além

⁸⁸ TATIT, L.; **Quantificações subjetivas: crônicas e críticas**; 2011; pp. 35-50.

disso, essas categorias recebem um investimento sêmico (Cf. Figura 15). (SOUZA; 2016: 56)

Figura 15 - “Quantificadores subjetivos de Zilberberg”

	S ₁	S ₂	S ₃	S ₄
ASCENDÊNCIA	só – extinção	– – restabelecimento	+ + recrudescimento	só + excesso
DESCENDÊNCIA	só + excesso	– + atenuação	+ – minimização	só – extinção

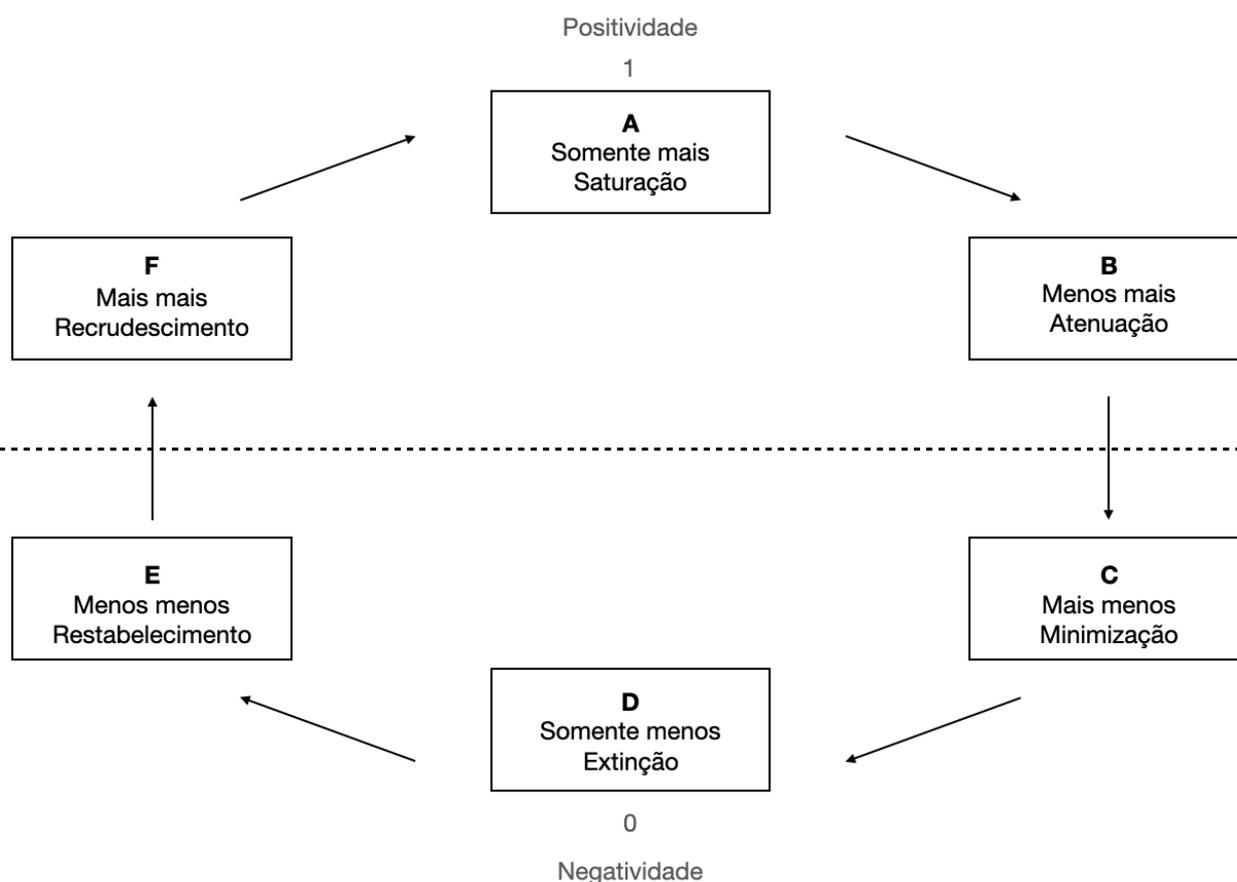
Fonte: reprodução (SOUZA; 2016: 56)

Entretanto, apesar da clareza exemplificativa, discordamos de que primeiro se estabelecem os limites das gradações - os "só mais" (saturação) e "só menos" (extinção). A nosso ver, os absolutos das grandezas encontram-se, para além da gradação, nas ultrapassagens realizadas pelo falante, em um espaço chamado por Beividas (2019; e no prelo) de superlativo ou supramolar. Para o pesquisador, as oposições molares (pequeno x grande; aberto x fechado) possuem certa prioridade em relação às oposições superlativas (minúsculo x imenso; hermético x escancarado). Desta feita, saturação e extinção parecem ser locais possíveis de se estabelecer apenas após uma definição anterior das grandezas molares.

Tendo nosso corpus como base, a apreciação alimentar manifestada nas avaliações do TripAdvisor parecem estabelecer bastante claramente o recrudescimento e a minimização como limites da gradação. Quando falamos de saturação, por outro lado, debruçamo-nos sobre casos extremos, de certa forma mais "raros", como o da fusão entre sujeito e objeto (uma das principais marcas da estesia). Nesse passo, quando se fala de extinção, tendo nosso corpus como base, parece sempre se relacionar a uma metamorfose que transforma objeto em abjeto. Essas "exceções" serão melhor aprofundadas nas subseções 3.2.1.2 (Saturação) e 3.2.1.5 (Extinção). Por enquanto, apenas para explicitar ainda mais as dificuldades relativas à instituição dos limites das gradações nos extremos absolutos, cabe mencionar a definição deles, dada por Tatit, como de regiões de curto-circuito teórico. Para tanto,

apresentamos também a nova roupagem (figura 16) dada pelo pesquisador à figura 15.

Figura 16 - Direções e segmentos tensivos⁸⁹



Fonte: TATIT (2019:229)

Nos campos extremos teríamos então, respectivamente, um acúmulo exclusivo de mais (em A) e um acúmulo exclusivo de menos (em D). A homologia com o princípio silábico indica, como já frisamos, que esses campos extremos não contribuem para a continuidade dos eventos e dos processos: a saturação por excesso de mais ou a extinção por excesso de menos provocam "pane" ou suspensão das forças. Podemos dizer, portanto, que a atenuação de A para B e o restabelecimento de D para E constituem modulações necessárias à manutenção da dinâmica tensiva. (TATIT; 2019: 230)

Ou seja, a saturação e a extinção configuram-se como caixas-pretas. Regiões que parecem catapultar as magnitudes sob suas influências para outras direções/

⁸⁹ Voltaremos inúmeras vezes a esta imagem. Entretanto, para evitar remissões a uma página bastante distante, todas as nossas amostras de classificação contemplam o mesmo diagrama.

dimensões⁹⁰. Tendo em vista o nosso corpus, somos levados a crer que a extinção, por exemplo, estabelece a direção do boicote, que é a sanção como programa de uso do enunciador (e não mais como programa de base, conforme a grande maioria de textos deste tipo); ou a direção do "quero meu dinheiro de volta", que parece deslocar tais discursos para uma esfera muito mais jurídica, da sanção negativa, conforme exposto em 2.3.

Dito isso, outro ponto importante de se salientar é que as quantificações subjetivas são mais visíveis naqueles textos em que uma apreciação é manifestada. Tal fato parece exigir que o ator da enunciação seja, no mínimo, um sincretismo entre sujeito de estado e destinador-sancionador. Para ilustrar a situação, Tatit vale-se de uma coluna escrita pelo historiador Jorge Coli que foi publicada em 7 de abril de 2008 no jornal *A Folha de São Paulo*. No texto jornalístico, o autor pondera que "Muita gente, no entanto, se irrita com ópera. As vozes lhes parecem poderosas demais e os sentimentos, excessivos." (TATIT; 2011: 37). Na análise, evidencia-se uma relação subjetiva em que o ator "muita gente" parece exigir uma atenuação do excesso.

Tanto o sentimento de falta quanto a impressão de excesso pressupõem a intermediação de um avaliador, o chamado destinador julgador, que, no interior de um quadro axiológico, estabelece seus parâmetros de medida, arbitrários do ângulo da exatidão científica, mas facilmente reconhecíveis como senso comum num determinado grupo sociocultural. No caso em exame, "muita gente" é o ator que ocupa a posição do avaliador cujo ponto de vista não coincide em nada com o do enunciador do texto. É o avaliador que identifica a ação antagonista e sustenta a resposta do sujeito no sentido de recompor as condições para superar a falta ou, como nesse caso particular, de moderar os efeitos do excesso. Muitas vezes, as funções de avaliador e de sujeito que reage são absorvidas pelo mesmo ator e isso contribui para que a resposta ao antissujeito seja mais rápida. (TATIT; 2011: 37)

Actancialmente falando, um sujeito que reage é um sujeito que responde a uma transformação em seu estado. Caso bastante óbvio, em nosso corpus, conforme já evidenciamos em 2.2. O ator da enunciação "usuário do TripAdvisor", entretanto, não parece preocupar-se em responder tonicamente ao antissujeito, mas apenas reportar a presença dele quando de casos de extinção e minimização.

⁹⁰ Em 3.2.1.2. Saturação e em 3.2.1.5. Extinção apresentaremos brevemente questões relativas à fusão de sujeito e objeto, que caracteriza certa estesia da saturação, e à transformação do objeto em abjeto, na extinção.

3.2.1. Exemplos de classificação

Selecionamos algumas avaliações para demonstrar a pertinência do sistema e da metalinguagem, desenvolvidos por Claude Zilberberg, na descrição de um universo de sentido. Reproduziremos avaliações da mesma forma como elas aparecem no TripAdvisor, inclusive para ressaltar a elasticidade da língua. Novas pesquisas poderão debruçar-se sobre a relação quantitativa estabelecida entre mensagens mais curtas e/ou mais longas. Como hipótese, parece plausível afirmar que no TripAdvisor as mensagens extensas são caracterizáveis como não-positivas. Somos levados a crer, assim, que haja um maior desejo coletivo no sentido de esclarecer ao máximo um problema, o que leva o ator da enunciação a utilizar formas discursivas muito mais longas do que nos casos de sanção positiva. Acreditamos, portanto, que as relações quantitativas estabelecíveis em relação ao tamanho dos textos seja fonte de ricas informações para pesquisas futuras.

Com relação aos exemplos de minimização e extinção, achamos pertinente a reprodução de novas formas metatextuais: a resposta do restaurante. Aparentemente, tais estruturas só estão disponíveis no caso de alta não-positividade discursiva. Mesmo que tal hipótese não venha a se confirmar, e que não seja possível traçar tal tipo de correlação comportamental, acreditamos que o exemplo seja ilustrativo para certa característica possível de desenvolvimento computacional junto a nosso modelo de inteligência artificial: a de oferecer resposta automatizada para os textos classificados com alta não-positividade. Tal fato ficará mais claro nas figuras 20 e 21, quando será possível perceber que os restaurantes já costumam praticar o envio de mensagens padronizadas em tais casos.

Nesta seção, entretanto, focaremos apenas na demonstração do poder heurístico da hipótese tensiva frente ao fenômeno do TripAdvisor. Demais depreendimentos computacionais serão desenvolvidos em pesquisas futuras. Destarte, e para definirmos grande parte de nossas classes, utilizaremos os glossários disponibilizados por Claude Zilberberg, e que parecem permanecer iguais, em *Eléments de grammaire tensive* (2006), *Des formes de vie aux valeurs* (2011) e *La structure tensive* (2012). Por questão de facilidade e disponibilidade (bem como de democratização de acesso e homogeneização de metalinguagem), daremos

preferência à versão da edição brasileira *Elementos de semiótica tensiva* ([2006] 2011), traduzida por Ivã Carlos Lopes, Luiz Tatit e Waldir Bevidas.

Todos os exemplos utilizados nesta seção foram retirados do perfil do restaurante paulistano D.O.M., de Alex Atala, que em 2012 ocupou o quarto lugar no ranking dos 50 melhores restaurantes do mundo, segundo a revista inglesa *Restaurant*⁹¹. Ainda hoje é a melhor posição já ocupada por um restaurante brasileiro na lista. Entretanto, a casa paulistana na última lista (2019) despencou para a 54ª posição. Já no TripAdvisor, o restaurante ocupa a 171ª posição, com 1.773 avaliações⁹².

3.2.1.1. Recrudescimento

Esta é a classe que, aparentemente, pelo menos no corpus utilizado para desenvolvermos nossa prova de conceito (confira 3.3 e apêndice), existe em maior quantidade entre as avaliações do TripAdvisor. Quando em 2.3 (Tipologia das sanções) apresentamos os inconvenientes causados pela implicação greimasiana não ser diferenciada entre tônica e átona, utilizamos o gradiente tensivo para demonstrar a solução proposta por Claude Zilberberg. No modelo desenvolvido pelo semioticista francês, entra em jogo o conceito de direção, que é estabelecido como o percurso que liga o 0 ao 1. Se tivermos em mente a solução de Tatit (figura 16⁹³), batizada como direções tensivas, não fica difícil perceber que existe uma linha separando a positividade e a negatividade. Falar de recrudescimento, assim, é falar de algum número que encontra-se entre o 0,5 e o 1 (mas muito mais próximo deste).

O recrudescimento é um dos dois analisadores (o outro é o restabelecimento) da ascendência tensiva percorrida de "0" a "1". Se interpretamos o restabelecimento como uma operação destinada a excluir, a subtrair um a um os menos, o recrudescimento se inscreve como a operação destinada a acrescentar um a um os mais. (ZILBERBERG; [2006] 2011: 272)

No TripAdvisor, os discursos difundidos parecem ser fortemente deste tipo. São textos em que, pouco a pouco, os atores vão distribuindo e enumerando os mais, as qualidades do restaurante em revisão. Vejamos um exemplo na figura 17, abaixo.

⁹¹ Disponível em: <https://www.theworlds50best.com/previous-list/2012>. Consulta: 27/11/2020.

⁹² Disponível em: https://www.tripadvisor.com.br/Restaurant_Review-g303631-d781071-Reviews-D_O_M_Gastronomia_Brasileira-Sao_Paulo_State_of_Sao_Paulo.html. Consulta: 27/11/2020.

⁹³ Também disponível na figura 17, abaixo.

Figura 17 - Avaliação do tipo recrudescimento



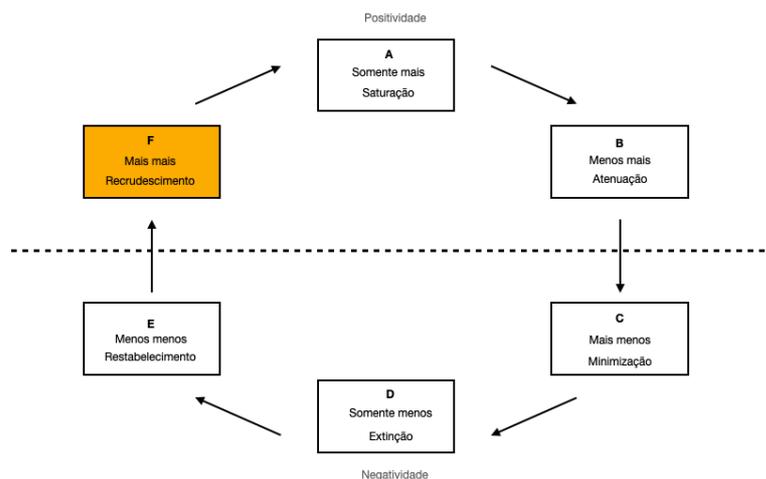
vanessaharumi
2 avaliações



Publicada em 31 de janeiro de 2020  via dispositivo móvel

Melhor experiência gastronômica

Sem sombra de dúvida, a melhor experiência gastronômica que tive na vida com o privilégio de ter ingredientes brasileiríssimos nos pratos. Sem duvida, o querido e muito atencioso Alex Atala é um gênio!



Fonte: adaptação a partir de reprodução/ TripAdvisor; e TATIT (2019)

No texto acima podemos perceber que as avaliações privilegiam uma relação entre destinador (usuário) e destinatário (restaurante) marcada por termos enunciativos como "(eu) tive" e "(o restaurante forneceu) a melhor experiência gastronômica". A vivência do ator da enunciação é descrita na forma das adições de mais reconhecidas por ele como positivas: "ingredientes brasileiríssimos" e "atencioso Alex Atala".

Se abandonarmos brevemente o exemplo fornecido pela avaliação do restaurante D.O.M. e buscarmos em nosso corpus algumas avaliações (confira 3.3 e o apêndice para mais informações) que foram anotadas como recrudescimento, podemos caracterizar melhor as informações fornecidas em *O Globo*⁹⁴. Num primeiro caso, o ator da enunciação privilegia o serviço e o figurativiza bastante fortemente: "Jantar de negócios com serviço de primeira dos garçons: João Paulo e Wagner e o Somellier Wellington"(sic). O texto não deixa dúvidas de que estamos em uma

⁹⁴ Em 2.1 apresentamos reportagem a respeito de pesquisa realizada pela UFRJ em que se afirma que o serviço é a principal categoria mencionada pelos consumidores (fato levantado por meio de n-gramas).

ascendência tensiva que evidencia o serviço em detrimento da comida. Mas isso não é regra. São bastante frequentes avaliações mistas, nas quais serviço e comida configuram-se com os mais que vão entrando: "Leonardo atendimento muito bom gente finíssima! Comida muito boa! Muito atenciosos! Parabéns!" (*sic*) Neste caso, temos também uma figurativização do tema do serviço representada pelo "garçom Leonardo", que mais do que um profissional competente para a realização do seu serviço, é configurado pelo ator como um ser humano de primeira, "gente finíssima". Entretanto, a comida também aparece como um mais, apesar de o tema ser figurativizado bem mais fracamente com um simples "Comida muito boa!", ao que o ator logo em seguida recrudescer ainda mais o serviço.

Tendo em vista tudo o que já dissemos sobre a elasticidade da língua, é complicado, no atual momento de pesquisa, afirmar que o tema do serviço seja mais frequente do que o tema do alimento. Principalmente quando nossas tabelas de n-gramas (Cf. tabelas 1-4) parecem evidenciar que, em relação ao serviço, há mais variações de formas de figurativizar o tema do que propriamente uma quantidade superior de figuras do serviço em detrimento de figuras da comida. A nosso ver, para sermos mais consequentes com a episteme semiótica, seria necessário o desenvolvimento de um "lematizador" de temas e figuras. Um programa capaz de agrupar as figuras primeiro por temas, propriamente dito, e depois dizer quais são as variedades de uma mesma lista de figurativizações⁹⁵.

Deixando de lado esse problema, o qual não é escopo do presente trabalho, e voltando para as avaliações, mencionamos que o texto que figurativiza o garçom Leonardo possui a operação de um recrudescimento duplo. O serviço é primeiro especificado na forma de um ator e depois generalizado para todos os demais responsáveis pelo "Muito atenciosos!" Esse ponto merece bastante atenção pois fala fundo sobre a teoria de que nos valem.

Após a categoria do recrudescimento, a "saturação" institui um lugar muito especial na teoria. Um bolsão, que é a expressão utilizada por Zilberberg em inúmeros textos. Nela, as coisas meio que saem do controle. A lógica cotidiana é abandonada e o ator experimenta uma vivência que muitos chamam de plenitude. Entretanto, chegar nesse lugar não é fácil e, uma vez nele, a permanência não é duradoura. A

⁹⁵ Algo muito próximo ao que foi feito por Pierre Maranda e apresentado em 1.2.

“saturação” se constitui (junto com a “extinção”) em lugares especiais da teoria. Tanto é, que para sairmos do recrudescimento e ingressarmos nessa zona obscura (ou *Twilight Zone*, se quisermos nos referir ao famoso seriado americano de ficção científica), não é tão simples.

Nem sempre o recrudescimento leva de maneira automática à saturação, pois, em princípio, podemos recrudescer o próprio recrudescimento (prática bastante utilizada nas criações poéticas e nas artes em geral, mas que pode ser verificada em frases simples como: “embora já fosse muito rico, nos últimos tempos aumentou ainda mais a sua fortuna”), assim como podemos minimizar infinitamente a minimização (como fazem as ciências moleculares e a nanotecnologia ou mesmo alguns relatos científicos: “embora considerado indivisível, o átomo foi fracionado em fins do século XIX”). Bem menos previsíveis, esses “excessos” acontecem com certa frequência e podem ser reconhecidos em diversos setores culturais. (TATIT, 2020: 12)

Conforme o excerto deixa claro, o recrudescimento recursivo é bem mais comum do que a saturação e está presente diariamente em nossas falas (algo bastante diferente da saturação). Ou seja, não importa o quanto o usuário enumere qualidade positivas de um restaurante, ele não sai, não avança necessariamente em direção à plenitude.

(...)em ascendência, passamos por implicação doxal do restabelecimento ao recrudescimento, como se passa do imperfectivo ao perfectivo com o intuito de amenizar uma insuficiência. Se a suficiência estiver realizada, é evidente que aquele que pretende ir além, estará atuando por concessão, desafio ou provocação; podemos dizer que, ao recrudescer o recrudescimento, esse alguém exerce o que poderíamos chamar de seu "direito imprescritível" à recursividade: a realização de um ultrapassamento. (ZILBERBERG, [2006] 2011: 273)

Ou seja, uma vez que a relação de conjunção entre sujeito e refeição é suficiente para o ator da enunciação, todos os /mais/ a mais que entram na cadeia parecem constituir premissa básica dos discursos. Quando a expectativa do sujeito não é muito alta e é atendida de acordo com a doxa (o senso comum), tudo o que está além do esperado ainda não parece ser capaz de instituir o inesperado, mas simplesmente reforçar a satisfação já de início anunciada - e cada vez mais reforçada.

3.2.1.2. Saturação

Para definir a classe do recrudescimento, nos vimos obrigados a adiantar a categoria da saturação - a qual inclusive comparamos com o título *Twilight Zone*, de uma série americana em que a cotidianidade era sempre rompida por acontecimentos surpreendentes e de outro mundo. Em linhas gerais, essa parece ser a melhor definição da categoria em que agora nos encontramos.

O recrudescimento, ou aumento de mais, pode levar a uma plenitude, típica dos finais de fábulas que “acabam bem”, mas pode igualmente ultrapassar esse nível eufórico e se converter numa espécie de “mais em demasia”, uma saturação, que torna urgente, ou inevitável, a diminuição dos elementos considerados excessivos.” (TATIT, 2011: 45)

No excerto, Tatit menciona a palavra plenitude. Antes de falarmos do termo, que acreditamos ser um dos mais importantes definidores da saturação, é preciso fazer um breve comentário. O semiótico brasileiro, no artigo em que escreveu aquelas palavras, está trabalhando com textos que atuam no nível da cotidianidade. Por isso ressalta que tão logo se chega ao “mais em demasia”, há a necessidade de uma atenuação (classe que veremos a seguir). Em “Quantificações subjetivas: crônicas e críticas”, quando menciona essa necessidade urgente de atenuação, Tatit está descrevendo discursos que costumam depreciar o gênero musical da ópera como sendo excessivo. Entretanto, não são poucos os textos que descrevem a saturação em si mesma, e não como uma etapa a ser vencida em prol do cotidiano. São os textos que trabalham com a estesia, a experiência estética.

Entre os bons exemplos, Greimas dedicou um livro inteiro (*De l'imperfection* ou *Da imperfeição*, em tradução portuguesa) para analisar o tema. Por lá, o fundador da semiótica discursiva, na primeira parte, “La fracture” (A fratura), debruça-se sobre cinco textos escritos por Julio Cortázar, Jun'ichirō Tanizaki, Italo Calvino, Michel Tournier e Rainer Maria Rilke. Em comum, todos eles retratam diferentes tipos de experiências com a estesia. Retomemos alguns dados pela eficaz síntese de Fiorin:

A experiência estética é um evento extraordinário enquadrado pela cotidianidade (1987:19), é uma surrealidade englobada pela realidade (1987:32). Nela o tempo pára, o espaço fixa-se (1987:15-16) e ocorre um sincretismo entre sujeito e objeto (1987:31), que estão disjuntos na temporalidade de todos os dias. Rasga-se o

parecer imperfeito e aparece a "nostalgia da perfeição", "oculta pela tela da imperfeição", que constitui a realidade cotidiana (1987:17). A estesia é o vislumbre do contínuo, da fusão anterior à discretização, perdida pela constituição da significação. (FIORIN, 1999: np)

Apesar desse contexto tão erudito, se não pouco usual, conforme o resumo de Fiorin para as principais características da experiência estética, não é preciso ser um grande artista para sentir tamanho fenômeno extraordinário. No TripAdvisor, não são poucos os usuários que relatam terem vivido fenômenos tão incríveis quanto os personagens da literatura.

Figura 18 - Avaliação do tipo saturação



Fonte: adaptação a partir de reprodução/ TripAdvisor; e TATIT (2019)

Na figura 18, vemos exemplificados todos os pontos sintetizados por Fiorin - apesar de o efeito de sentido da estesia nas avaliações de restaurante ser um pouco menos poético e bem mais marqueteiro. Uma primeira diferença nos textos de saturação, em relação aos de recrudescimento está na debreagem de um observador, figurativizado no texto pela expressão "Se você vai ao DOM" (além de inúmeras outras figuras da segunda pessoa, o tu, que ninguém sabe muito bem quem é). A impressão que temos é a de que o ator da enunciação entrou em uma conjunção tão grande com o objeto refeição, ficou tão encantado com a experiência, que está até falando sozinho. É isso o que caracterizamos como plenitude, esse estado de fusão

entre sujeito e objeto. Apesar de o fenômeno ser muito parecido com aquele explicitado pelos grandes escritores da literatura mundial, no TripAdvisor ele normalmente tem um quê de panfletagem. O ator da enunciação parece tomar partido do restaurante e começa a falar por ele. Somos levados a imaginar, inclusive, que estamos perante uma publicidade.

Essa fusão é, na verdade, uma mudança de plano enunciativo. O sujeito passa do plano da enunciação enunciada para o do enunciado enunciado. Em *A rosa púrpura do Cairo*, de Woody Allen, a mulher, maltratada pelo marido brutal, refugia-se no cinema para esquecer as agruras de sua vida triste. Sua fusão com o objeto fílmico é figurativizada pela entrada na ação do filme, para viver uma história de amor, uma vida cheia de aventuras, com o galã do cinema. Nessa outra dimensão enunciativa, o sujeito deixa a realidade da existência, para viver, durante o tempo da experiência estética, uma surrealidade, uma segunda vida. Esses exemplos mostram a fecundidade da questão teórica das pré-condições da significação, para a compreensão de certos tipos de textos." (FIORIN, 1999: np)

Nosso ator da enunciação, feito a mulher de *A rosa púrpura do Cairo*, quer fugir da realidade. Quer "comer formiga e achar bom", quer pagar "bem caro", mas achar que "vale cada centavo". Ele não vai ao restaurante buscando uma alimentação ou um jantar casual. Ele quer um "menu maximus com harmonização". Ele quer ser surpreendido. Os traços mambembes, que mencionamos e aludimos à publicidade, parecem depreender-se de uma falta de saber-fazer que empobrecem a configuração discursiva⁹⁶ da estesia. Apesar de buscar o inesperado, de querer fugir do cotidiano, o ator da enunciação da figura 18 passa a "receita do bolo" para alguém que queira repetir a experiência estética. E aí a gente saiu da estese para cair na "reprodução artística". Coisa que talvez Walter Benjamin, Adorno entre outros, poderão ajudar a esclarecer futuramente.

3.2.1.3 Atenuação

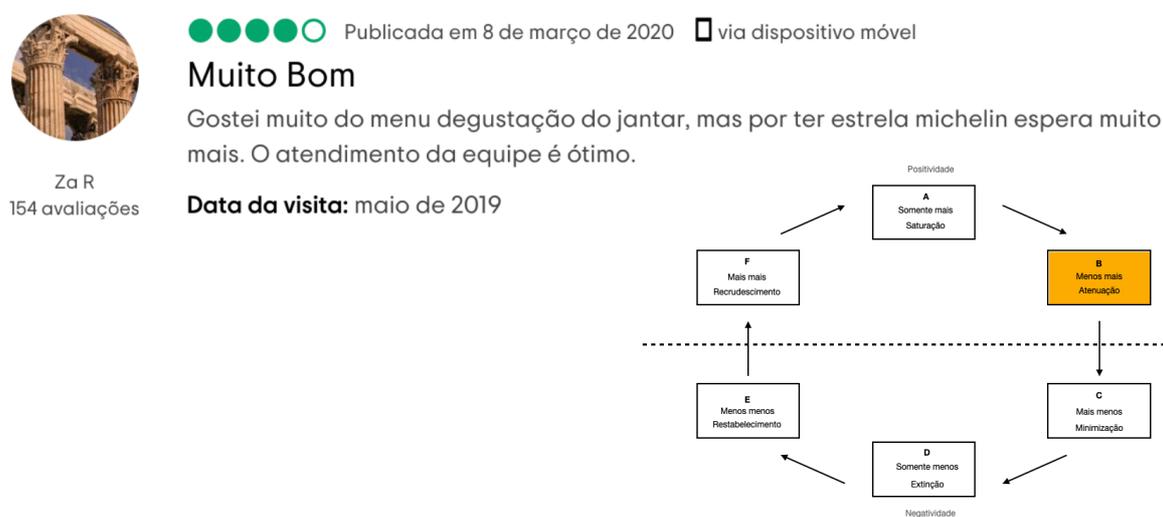
Muito mais próxima de nós, e já mencionada algumas vezes nas classes acima, a atenuação guarda relação íntima com retórica. Nos casos em que a saturação é vista como excessiva, fato descrito por Tatit a respeito do sentimento que muitas

⁹⁶ Pessoa de Barros apresenta bom debate sobre o tema no artigo "Problemas de enunciação" (in *Cruzeiro Semiótico*; 1987; pp 68-75. Disponível em: <https://felsemiotica.com/descargas/cruzeirosemiotico6.pdf>. Consulta: 08/02/2021.

peças relatam em relação à Ópera, um sentimento de redução do excesso faz-se necessário para instituir o curso natural das coisas.

A atenuação é um dos analisadores da minimização da descendência tensiva. Seja a série descendente simples de direção $[1 \rightarrow 0]$, a atenuação da tonicidade, por exemplo, associa o sobrecontrário tônico [S1] e o subcontrário tônico [S2]; a atenuação supõe a "adição" de um menos. A atenuação é o correlato a quo da minimização. (ZILBERBERG; [2006] 2011: 239)

Figura 19 - Avaliação do tipo atenuação



Fonte: adaptação a partir de reprodução/ TripAdvisor; e TATIT (2019)

Criador da hipótese tensiva, Zilberberg não perde muito tempo descrevendo esse termo cuja simplicidade é bastante enganadora. No excerto acima, é preciso atenção para conceitos semioticamente importantes. Quando o autor menciona o correlato *a quo*, por exemplo, ele quer dizer que a minimização supõe uma atenuação anterior. Ou seja, a direção descendente inicia-se pela atenuação. Entretanto, é preciso cuidado para não simplificarmos o conceito e o colocarmos diretamente no polo da negatividade. Tatit, por exemplo, por meio da figura 16⁹⁷, diferencia duas atenuações: uma positiva e uma negativa.

Do campo B ao campo C, teremos duas resoluções tensivas: de B até a linha pontilhada horizontal, verificamos perda de positividade, ou seja, menos mais; da

⁹⁷ Cf. Figura 19, acima, para evitar remissão à página 80.

linha horizontal ao campo C constatamos "ganho" de negatividade, ou seja, mais menos. (TATIT; 2019: 230)⁹⁸

3.2.1.4. Minimização

A minimização é o contrário do recrudescimento. Se a cadeia de /mais mais/ já foi apresentada em 3.2.1.1, agora é a hora de demonstrar a heurística da metalinguagem zilberberguiana junto à cadeia de /mais menos/. A presente classe aparece em nosso corpus fortalecendo a descendência tensiva [1 → 0] iniciada pela atenuação.

A minimização tem como marca o intervalo [S3 + S4], ou seja, é constituída pelo subcontrário átono [S3] e pelo sobrecontrário átono [S4]. Do ponto de vista actancial, a intensidade, após a adoção da tese de G. Deleuze, dirige-se - caso não seja introduzido um contraprograma eficaz determinado pela modalidade do prevenir - para a sua própria anulação. Do ponto de vista valencial, pelo fato de suceder à atenuação, a minimização só contém menos, ou seja, mais nada. A minimização é o correlato ad quem da atenuação. (ZILBERBERG; [2006] 2011: 268)

A explicação de Zilberberg não poderia ser mais esclarecedora e parece dar conta do fenômeno resposta do restaurante. Conforme falamos no início desta seção, nos casos de minimização e extinção, parece haver um certo consenso entre parte dos donos dos estabelecimentos de que seja necessário a entrada de um contraprograma, de uma resposta ao usuário do TripAdvisor. Muitas ilações são possíveis, inclusive se tivermos em mente as palavras de Kotler, apresentadas em 2.1⁹⁹. Entretanto, no atual momento de pesquisa, é impossível fornecer, quantitativamente, a correlação entre minimização e extinção e as respostas dos restaurantes. Ainda assim, a figura 20 logo abaixo evidencia tal fenômeno.

⁹⁸ Para evitar a remissão até a figura 16, a figura 19 (bem como todas as demais, com exemplos de classificação) possuem o diagrama das direções tensivas. Na roda, separando os campos da positividade e da negatividade, é possível visualizar uma linha pontilhada. É sobre isso que Tatit fala na citação.

⁹⁹ Para o "guru do marketing, os cidadãos da internet podem facilmente ser convertidos em advogados da marca (para o bem e para o mal). Responder a uma crítica não positiva, portanto, parece ser uma forma de os restaurantes não perderem um cliente e, mais do que isso, atenuar uma má impressão que pode se tornar propaganda negativa do estabelecimento.

Figura 20 - Avaliação do tipo minimização¹⁰⁰

Danuse
Sorocaba, SP
19 6

Publicada em 6 de julho de 2019

Desapontamento e alerta!

Menu degustação de 20 anos da casa começa morno e atinge seu auge no pirarucu com paçoca...poderia ter continuado nas variações de mandioca. O desapontamento veio quando, ao recebermos a lista dos pratos no final do menu, percebemos que o prato veio incompleto - um erro amador que não se espera de uma casa com o renome do DOM. Quando ligamos para informar o ocorrido, percebemos descaso. Se restava alguma dúvida quanto à queda de 50 posições nos últimos anos do ranking 50 Best, está esclarecida.

Mostrar menos

Data da visita: julho de 2019

Custo-benefício

Atendimento
Comida

Ver todas as 13 avaliações de Danuse sobre São Paulo

Peça informações para Danuse sobre D.O.M. Gastronomia Brasileira.

3 Obrigado, Danuse

Esta avaliação representa a opinião subjetiva de um membro do programa TripAdvisor e não da TripAdvisor LLC.

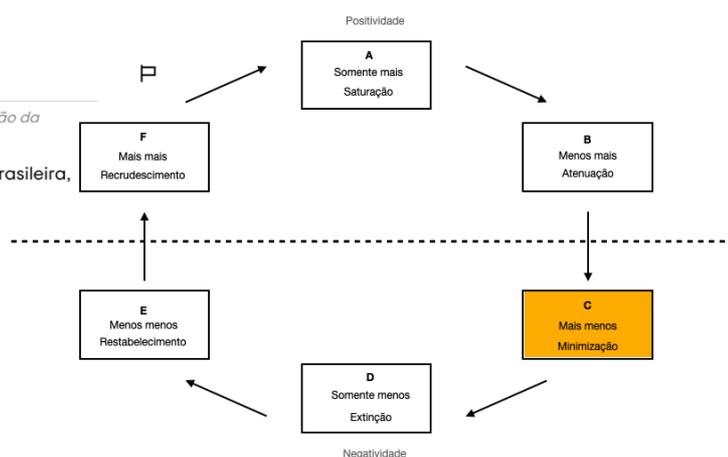
GabriellaP1492, Gerente de relações públicas do(a) D.O.M. Gastronomia Brasileira, respondeu a esta avaliação

Respondida em 10 de julho de 2019

Olá, Danuse,

Pode, por favor, nos passar um telefone para contato pelo e-mail comunicacao@domrestaurante.com.br
Muito obrigado.

D.O.M. Restaurante



Fonte: adaptação a partir de reprodução/ TripAdvisor; e TATIT (2019)

Nesta categoria, ainda fica clara a realização da relação de junção entre S2 e o objeto refeição (O). Entretanto, de acordo com o quadro de valores do usuário do TripAdvisor, vemos que houve uma quebra de contrato: “percebemos que o prato veio incompleto”. Depreende-se do texto na figura 20 que o impresso, com a sequência de pratos servidos no menu degustação, não estava disponível desde o começo. Foi apenas ao chegar em casa que o usuário percebeu ter pago por algo que não comeu. Feito pedir uma feijoada completa, mas o restaurante esquecer a bisteca ou o torresmo. Entretanto, de acordo com o texto, o esquecimento de um prato na sequência do menu já se configurava dentro de uma descendência tensiva.

Segundo o usuário, a refeição já começara morna, sem empolgação. Tal experiência abaixo das expectativas pode ser considerada como o primeiro /menos/. Após este primeiro /menos/, ainda somaram-se outros dois: o esquecimento dos pratos e o

¹⁰⁰ Ressalte-se a diferença para as figuras anteriores: a resposta do restaurante.

descaso com a reclamação. “Quando ligamos para informar o ocorrido, percebemos descaso”.

Foram três /menos/ que representam uma verdadeira recursão. Principalmente quando o usuário sintetiza que todas essas quebras de contrato não eram todavia inesperadas. “Se restava alguma dúvida quanto à queda de 50 posições nos últimos anos do ranking 50 Best, está esclarecida.”¹⁰¹

3.2.1.5. Extinção

Da mesma forma como no caso da saturação, nesta classe encontramos discursos de outro nível. No lugar da conformação tradicional entre sujeito e objeto, por aqui, um novo actante entra em cena: o abjeto.

Se nos voltamos à ordem remissiva, reencontramos o abjeto - termo que emprestamos de J. Kristeva. À primeira vista, essa categorização parece oferecer um certo grau de evidência, particularmente no nível narrativo: os enunciados de não-conjunção com o objeto podem ser considerados com explicitações, catálises, discursivizações do abjeto. (ZILBERBERG; [1988] 2006: 141-142)

Para Zilberberg, o abjeto está sempre relacionado com a não-conjunção, o que, em nosso caso, pode acontecer de inúmeras formas. Em discurso, da mesma forma que a posição de fusão entre sujeito e objeto aparece como saturação, quando o sujeito quer se afastar da refeição, quando o “não querer se alimentar” configura-se como a ação principal do usuário, deparamos esse actante novo. Em termos semióticos, podemos traduzir esse fato pela fórmula canônica /não-querer-ser/, ou seja, o sujeito de estado (S2) não quer entrar em conjunção com o objeto (O). Os motivos (ou seja, os revestimentos discursivos) podem ser os mais variados. Entretanto, todos parecem estar sob o guarda-chuva da repulsa.

¹⁰¹ Conforme mencionamos em 3.2.1, o restaurante D.O.M. já ocupou a 4ª posição (2012) no ranking mencionado pelo usuário. Em 2019, caiu para a 54ª posição.

Figura 21 - Avaliação do tipo extinção¹⁰²

Mauricio Z
2 11

Publicada em 1 de fevereiro de 2020

Decepção!

Eu e minha esposa tentamos fazer uma reserva no DOM, e o e-mail que chegou até nós dizia que não tinha vagas para nosso horário e dia, então mandei um e-mail respondendo que eu não tinha interesse pra outras datas que me disponibilizaram, pois eu queria ir no dia do aniversário dela (30 de janeiro) agora acabei de ver meu cartao, e veio uma cobrança de 200 reais por não ter comparecido! Fui verificar meus e-mails, e do nada apareceu que tinha surgido uma vaga! Mesmo que eu já tinha mandado um e-mail desistindo de qqr dia! Sacanagem master! Agora vou gastar 200 reais por não ter ido, e sem saber que havia liberado uma mesa pra mim! Pq eu já tinha desistido! Nunca mais vi meus emails até pq estou em SP em viagem! Dom! Vá tomar no c...

Mostrar menos

Data da visita: fevereiro de 2020

Peça informações para Mauricio Z sobre D.O.M. Gastronomia Brasileira.

9 Obrigado, Mauricio Z

Esta avaliação representa a opinião subjetiva de um membro do programa TripAdvisor e não da TripAdvisor LLC.

NathaliaDOM, Gerente de relações públicas do(a) D.O.M. Gastronomia Brasileira, respondeu a esta avaliação

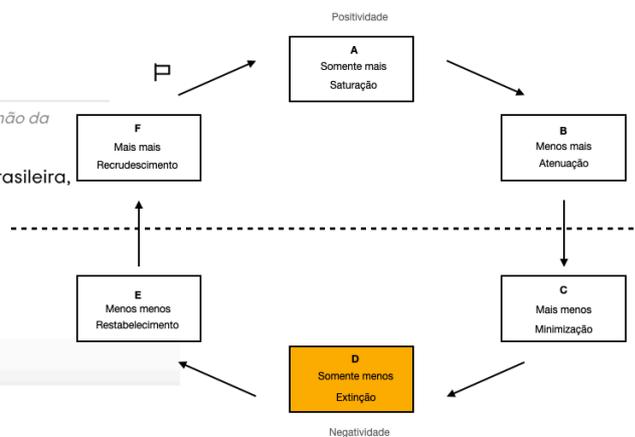
Respondida em 5 de fevereiro de 2020

Olá Mauricio,

Você poderia entrar em contato conosco através do e-mail comunicacao@domrestaurante.com.br, por favor?

Atenciosamente,

Equipe D.OM..



Fonte: adaptação a partir de reprodução/ TripAdvisor; e TATIT (2019)

Sendo assim, basta observarmos a figura 21 para compreendermos como se dá a "construção do abjeto". No texto, o ator da enunciação detalha em minúcias a transformação do objeto, que passou da meta máxima, figurativizada na tentativa da comemoração do aniversário da esposa em "grande estilo", a um caso típico de discursos do ressarcimento, do quero o meu dinheiro de volta.

(...)essas mesmas relações, subjetais e objetais, podem ser regidas pelo fazer remissivo, de modo que teríamos então, de um lado, rupturas polêmicas entre sujeito e antissujeito e, de outro, descontinuidade entre sujeito e objeto, o que pode fazer desse último um abjeto. (TATIT; 2019: 95)

Tatit caracteriza bastante bem a relação entre sujeito e abjeto como uma reversão no sentido 'correto' realizada pelo fazer remissivo. Neste caso, *grosso modo*, parece homologar a instituição da direção descendente com a remissividade. Tudo o que o ator não quer mais é entrar em conjunção. Ele desenvolveu um verdadeiro asco pelo

¹⁰² Na comparação com a figura anterior, percebe-se a facilidade (necessidade?) de se produzir respostas automatizadas nestes casos de alta não-positividade.

restaurante, figurativizado no xingamento apenas atualizado: “Dom! Vá tomar no c...”

3.2.1.6. Restabelecimento

Se na extinção nós teríamos como que a classe máxima da remissividade, iniciada pela atenuação, na classe do restabelecimento parece haver um germen de ascendência ou já a direção propriamente dita. Para Tatit, restabelecimento e atenuação possuem propriedades próprias. E ele utiliza a figura 16¹⁰³ para descrever tal fenômeno.

Do campo B ao campo C, teremos duas resoluções tensivas: de B até a linha pontilhada horizontal, verificamos perda de positividade, ou seja, menos mais; da linha horizontal ao campo c, constatamos ganho de negatividade, ou seja, mais menos. Do Campo E ao campo F, teremos as resoluções inversas: de E à linha horizontal, a perda de negatividade, ou seja, menos menos; da linha horizontal ao campo F, depreendemos ganho de positividade, ou seja, mais mais. (TATIT; 2019: 230)

Em nosso corpus, até o presente momento de pesquisa, ainda não encontramos exemplares discursivos do tipo restabelecimento. Entretanto, com base nos exemplos apresentados sobre as minimizações e extinções, é possível simularmos, com objetivos puramente descritivos, possíveis discursos em que tal classe se manifesta.

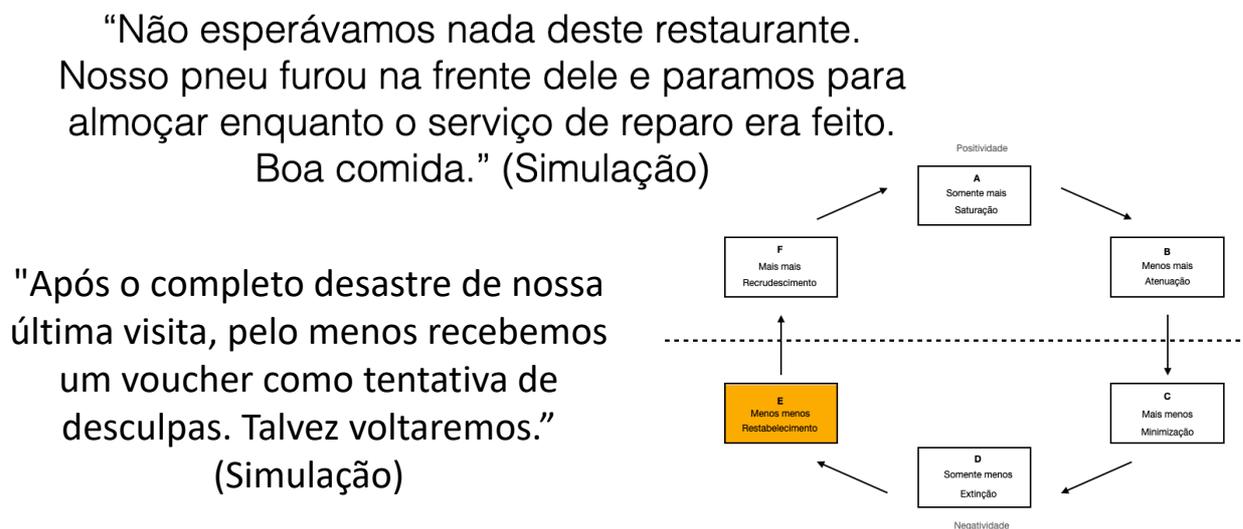
Na última vez que estivemos aqui, o garçom nos tratou muito mal, o camarão veio frio e a conta totalmente errada. Após reclamarmos que nunca mais voltaríamos, fomos surpreendidos por uma mensagem super simpática do gerente Amaury, que entrou em contato e foi super atencioso, se desculpando por todo o ocorrido e oferecendo um voucher que utilizamos hoje. Tudo estava impecável. (SIMULAÇÃO)

Acima, tentamos criar um exemplo de enunciado totalmente pertinente à plataforma TripAdvisor. Entretanto, no lugar de um simples restabelecimento, temos a passagem da extinção "nunca mais voltaríamos" para o recrudescimento "Tudo estava impecável". A velocidade com a qual o ator da enunciação passa da pura negatividade para a pura positividade é tremenda. Tal tipo de fenômeno parece bastante mais fácil de encontrarmos, basta atenção para as marcas de experiências

¹⁰³ As figuras 21 (acima) e 22 (abaixo) evitam a necessidade de remissão.

passadas no texto. Entretanto, não seria difícil imaginar outros casos, em que o usuário sairia da zona negativa pouco a pouco (Cf. figura 22, logo abaixo).

Figura 22 - Simulações de avaliação do tipo restabelecimento



Fonte: adaptação a partir de simulações e de TATIT (2019)

Esperamos que, após a sequência das duas classificações anteriores (minimização e extinção) e as simulações acima, seja mais fácil perceber as reversões impostas à refeição, que ora se configura como objeto e em outras como abjeto. Acreditamos que a baixa velocidade, a passagem desacelerada da extinção para o restabelecimento, torne a existência da classe rara. Nos parece que ninguém "perde muito tempo" para dizer que talvez volte ao restaurante. Em contrapartida, os extremos, "retornarei com certeza" ou "não volto nunca mais", são bastante representativos em nosso corpus. Tal fato parece denunciar um certo desinteresse cultural pelo desenvolvimento discursivo do restabelecimento. Como se ele não fosse digno de nota. Entretanto, mais pesquisas (e classificações) são necessárias para se confirmar esse fato.

3.3. Prova de conceito com rede neural

Segundo o dicionário Oxford, o termo prova de conceito ('proof of concept' ou simplesmente PoC) é uma "evidência que demonstra que uma proposta de negócio, ideia de design, entre outros, irá funcionar, usualmente, com base em um

experimento ou projeto piloto"¹⁰⁴. O termo diz respeito, bastante usualmente, a uma atividade que desobriga a realização completa de um projeto, pois oferece evidências suficientes da factibilidade da ideia. No atual estágio de nossa pesquisa, infelizmente não houve tempo para a total implementação de um modelo de inteligência artificial 100% funcional. Por conta disso, e para demonstrar a viabilidade de se utilizar a semiótica discursiva na representação de conhecimento, apresentaremos um sistema que ainda demanda otimizações. Mesmo assim, todos os códigos utilizados (e que serão apresentados como apêndices a este trabalho) estão bastante próximos ao estado da arte e podem ser utilizados em qualquer corpus anotado com o objetivo de desenvolver uma atividade de classificação automática com *deep learning*.

Para obtermos nosso corpus, foi utilizado um *script* de raspagem de dados (*web scraper*) por meio do qual coletamos 45 mil avaliações dos 31 restaurantes mais bem avaliados da cidade de São Paulo. Tal atividade aconteceu entre os dias 11 e 14 de fevereiro de 2020 e diz respeito a avaliações escritas no período de 2016 e 2020.

A partir dessa base de dados, selecionamos dois restaurantes tendo em conta a possibilidade de representarem o fenômeno o mais abrangentemente possível. Para tanto, optamos selecionar as avaliações da unidade Anália Franco da rede Coco Bambu, especializada em pescados e frutos do mar. Junto a ela, somamos também os textos referentes ao bar e restaurante Seen, localizado na cobertura do hotel Tivoli Mofarrej, na região da Avenida Paulista.

Desta forma, intentamos reunir informações a respeito de diferentes regiões da cidade de São Paulo (zona leste e centro) e sobre regiões que oferecessem ticket médio diferente. De acordo com a plataforma TripAdvisor, o Coco Bambu é representado por dois cifrões (\$\$) ao passo que o Seen por quatro (\$\$\$\$). Não há muito consenso a respeito da legenda para tal representação gráfica. Entretanto, alguns usuários propõem que sejam lidos como: \$\$, preços médios; \$\$\$\$, preços bastante elevados.

¹⁰⁴ Tradução nossa para: "evidence that shows that a business proposal, design idea, etc. will work, usually based on an experiment or a pilot project". Disponível em: <https://www.oxfordlearnersdictionaries.com/us/definition/english/proof-of-concept?q=proof+of+concept>. Consulta: 22/01/2021

O método de aproximação não se mostrou muito satisfatório, uma vez que a classificação aleatória das avaliações demonstrou haver um excesso de textos do tipo “recrudescimento” em detrimento de todos os outros. Tal fato causou sobreajuste (overfitting, em inglês, é o termo mais conhecido na área), que em aprendizado de máquina corresponde a quando um modelo não está aprendendo a identificar padrões, mas sim decorando os dados apresentados ou prevendo, desproporcionalmente, mais uma categoria do que outra. Demonstra baixa capacidade de generalização.

Em apêndice a este trabalho, disponibilizamos um notebook da implementação do código realizado na plataforma Google Colaboratory, que nos fornece a capacidade de realizar o treinamento da inteligência via disponibilização de GPU na nuvem. Tal fato, além de deixar nossos computadores desimpedidos para o uso durante a atividade de processamento, ainda realiza a performance em baixíssimo tempo, o que nos poupa muitas horas de trabalho.

Antes de apresentarmos nossos resultados e os processos utilizados, é preciso fazer uma breve exposição sobre o conceito de neurônio e de rede neural. Após isto, poderemos demonstrar a atividade de uma rede do tipo AWD-LSTM¹⁰⁵. Devido ao escopo deste trabalho, não poderemos nos estender muito nos debates por trás do funcionamento das redes. Nosso interesse é o de apresentar rudimentos sobre a área para incentivar que novas pesquisas sejam desenvolvidas.

O neurônio McCulloch-Pitts (McCulloch e Pitts, 1943) foi um dos primeiros modelos de função cerebral. Este modelo linear pode reconhecer duas categorias diferentes de entradas, testando se $f(x, w)$ é positivo ou negativo. Obviamente, para que o modelo corresponda à definição desejada das categorias, os pesos precisam ser definidos corretamente. Esses pesos podem ser definidos pelo operador humano. Na década de 1950, o perceptron (Rosenblatt, 1958, 1962) se tornou o primeiro modelo

¹⁰⁵ Utilizamos tal tipo de arquitetura devido a certa “coerção” da plataforma *Fast.AI*, que a considera como o estado da arte atual. Não nos prenderemos em detalhes sobre a utilização de tal arquitetura. Para aqueles que queiram entender mais as inovações de tal rede recorrente otimizada, recomendamos a leitura do artigo (MERITY; KESHAR; SOCHER; 2017)

que poderia aprender os pesos definindo as categorias dados exemplos de entradas de cada categoria.¹⁰⁶ (GOODFELLOW e BENGIO; 2016: 15)

Não são poucos os pesquisadores que reconhecem que a inspiração para a criação das redes neurais tenha sido pautada inicialmente (e por muito tempo) pelas neurociências. Entretanto, autores como Goodfellow e Bengio (2016) são bastante responsáveis quando afirmam que a capacidade de a computação, hoje em dia, de reproduzir a atividade neural de forma realista esteja bastante longe da realidade.

(...)embora a neurociência tenha inspirado com sucesso várias arquiteturas de rede neural, ainda não sabemos o suficiente sobre o aprendizado biológico para que a neurociência possa oferecer mais orientações para os algoritmos de aprendizado que utilizamos no treinamento de tais arquiteturas.¹⁰⁷ (GOODFELLOW e BENGIO; 2016: 16)

A principal contribuição das neurociências para as ciências da computação, segundo Goodfellow e Bengio foi a proposta de uma arquitetura em que as unidades computacionais atuam umas em cima das outras. “Cada unidade se assemelha a um neurônio no sentido de que recebe dados de muitas outras unidades e calcula seu próprio valor de ativação.”¹⁰⁸ (GOODFELLOW e BENGIO; 2016: 169) Segundo BENGFORT; BILBRO e OJEDA existem dois tipos de rede.

Em uma rede *feedforward*, os sinais viajam da camada de entrada para a camada de saída em uma única direção. Em arquiteturas mais complexas, como redes recorrentes e recursivas, a distribuição do sinal pode se combinar ou se repetir entre os nós de uma camada.¹⁰⁹ (2018: 275)

¹⁰⁶ Tradução nossa para: “The McCulloch-Pitts Neuron (McCulloch and Pitts, 1943) was an early model of brain function. This linear model could recognize two different categories of inputs by testing whether $f(x, w)$ is positive or negative. Of course, for the model to correspond to the desired definition of the categories, the weights needed to be set correctly. These weights could be set by the human operator. In the 1950s, the perceptron (Rosenblatt, 1958, 1962) became the first model that could learn the weights defining the categories given examples of inputs from each category.”

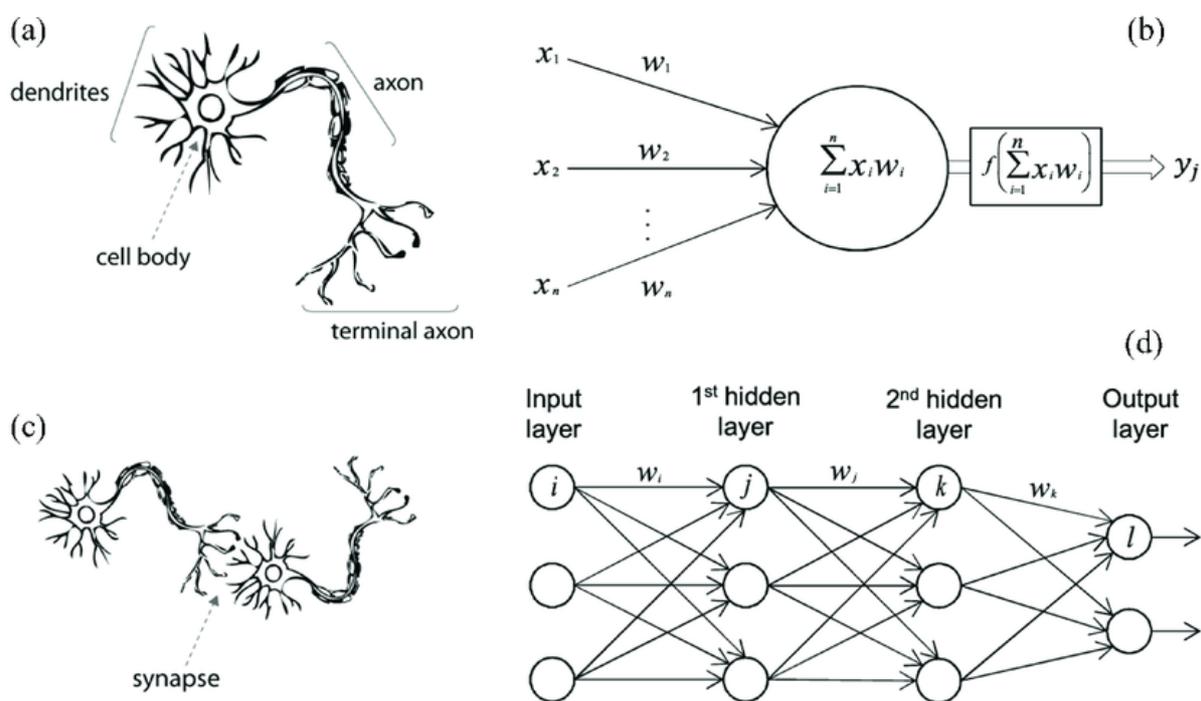
¹⁰⁷ Tradução nossa para: “(...)while neuroscience has successfully inspired several neural network architectures, we do not yet know enough about biological learning for neuroscience to offer much guidance for the learning algorithms we use to train these architectures.

¹⁰⁸ Tradução nossa para: “Each unit resembles a neuron in the sense that it receives input from many other units and computes its own activation value.”

¹⁰⁹ Tradução nossa para: “In a feedforward network, signals travel from the input to the output layer in a single direction. In more complex architectures like recurrent and recursive networks, signal buffering can combine or recur between the nodes within a layer.”

Na figura 23 temos uma comparação entre redes biológicas e artificiais. Após todo o exposto até aqui, é importante entender a abstração arquiteto-computacional realizada pelos principais pesquisadores da área. Em (a) temos uma representação de neurônio biológico e em (b) um artificial. É interessante notar que a leitura computacional abstrai dos dendritos e do corpo celular a recepção dos *inputs* (em b são os x_n) e o processamento como uma somatória. Pesos (w_n) de recepção são propostos individualmente para cada *input* antes de serem processados. Já em (c) e (d) temos representações da interação entre tais unidades.

Figura 23 - Comparação neurológica entre modelos biológicos e artificiais



Fonte: (MALTAROLLO; HONÓRIO e SILVA; 2013: 205)

Os dendritos e os terminais axônicos dão conta de representar aquilo que GOODFELLOW e BENGIO (2016) chamam de “função vetor-para-escalar”. Ou seja, no lugar de o neurônio receber um *input*, realizar um processamento e oferecer um output para o pesquisador, todo esse processo começa e termina dentro de uma rede com várias unidades computacionais.

A ideia básica de ter muitas unidades computacionais que se tornam inteligentes apenas por meio de suas interações entre si é inspirada no cérebro. O Neocognitron

(Fukushima, 1980) introduziu um poderoso modelo de arquitetura para processamento de imagens que foi inspirado na estrutura do sistema visual dos mamíferos e mais tarde se tornou a base para a moderna rede convolucional¹¹⁰ (GOODFELLOW e BENGIO; 2016: 16)

Para explorar mais essas relações entre cérebro biológico e rede neural artificial, GOODFELLOW e BENGIO utilizam o conceito de representação distribuída. A importância de que redes sejam criadas de tal forma diz respeito, muito fortemente, a otimização e economia processual.

Por exemplo, suponha que temos um sistema de visão que pode reconhecer carros, caminhões e pássaros e esses objetos podem ser vermelhos, verdes ou azuis. Uma maneira de representar essas entradas seria ter um neurônio separado ou unidade oculta que se ativa para cada uma das nove combinações possíveis: caminhão vermelho, carro vermelho, pássaro vermelho, caminhão verde e assim por diante. Isso requer nove neurônios diferentes, e cada neurônio deve aprender independentemente o conceito de cor e identidade de objeto. Uma maneira de melhorar essa situação é usar uma representação distribuída, com três neurônios descrevendo a cor e três neurônios descrevendo a identidade do objeto. Isso requer apenas seis neurônios no total, em vez de nove, e o neurônio que descreve a vermelhidão é capaz de aprender sobre a vermelhidão por meio de imagens de carros, caminhões e pássaros, não apenas de imagens de uma categoria específica de objetos.¹¹¹ (GOODFELLOW e BENGIO; 2016: 17)

Segundo BENGIO; BILBRO e OJEDA (2018), redes neurais não são nada novas, podendo inclusive ser consideradas como metodologias “old school” de inteligência artificial que não deram muito certo no começo. Entretanto, nas últimas duas décadas, devido ao aumento das capacidades de processamento via GPUs, da computação distribuída, das otimizações nas taxas de aprendizado e no código

¹¹⁰ Tradução nossa para: “The basic idea of having many computational units that become intelligent only via their interactions with each other is inspired by the brain. The Neocognitron (Fukushima, 1980) introduced a powerful model architecture for processing images that was inspired by the structure of the mammalian visual system and later became the basis for the modern convolutional network (LeCun et al., 1998b)”

¹¹¹ Tradução nossa para: “For example, suppose we have a vision system that can recognize cars, trucks, and birds and these objects can each be red, green, or blue. One way of representing these inputs would be to have a separate neuron or hidden unit that activates for each of the nine possible combinations: red truck, red car, red bird, green truck, and so on. This requires nine different neurons, and each neuron must independently learn the concept of color and object identity. One way to improve on this situation is to use a distributed representation, with three neurons describing the color and three neurons describing the object identity. This requires only six neurons total instead of nine, and the neuron describing redness is able to learn about redness from images of cars, trucks and birds, not only from images of one specific category of objects.”

aberto de bibliotecas Python como Keras, TensorFlow e PyTorch o cenário vem mudando.

(...)não apenas mais práticas, as redes neurais também prometem ganhos de desempenho não triviais em relação aos modelos tradicionais. Isso ocorre porque, ao contrário dos modelos tradicionais, que enfrentam platôs de desempenho mesmo quando mais dados se tornam disponíveis, os modelos neurais continuam a melhorar¹¹². (BENGFORT; BILBRO e OJEDA; 2018: 274)

Após tais explicações, podemos apresentar nossa primeira implementação de rede neural artificial. Nela utilizaremos uma versão de rede recorrente (AWD-LSTM, conforme já comentamos acima) para processar nossas avaliações. Antes disso, é importante apresentar nosso corpus. Conforme a figura 24 (abaixo), é possível ver que ele possui 2602 linhas. Ou seja, aquele é o montante de avaliações realizadas no TripAdvisor que foram classificadas e serão utilizadas no treinamento de nosso modelo de inteligência artificial. Também podemos observar que, além da coluna com os textos, outras duas colunas (com dados no formato de números inteiros) dão conta de duas possibilidades de classificação: direção tensiva e segmento tensivo¹¹³.

Figura 24 - Metadados do corpus de treinamento de inteligência artificial

```
RangeIndex: 2602 entries, 0 to 2601
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  -
0   review_body     2602 non-null    object
1   segmento        2602 non-null    int64
2   direção         2602 non-null    int64
dtypes: int64(2), object(1)
memory usage: 61.1+ KB
```

Fonte: produção própria a partir do Colaboratory e da linguagem Python de programação

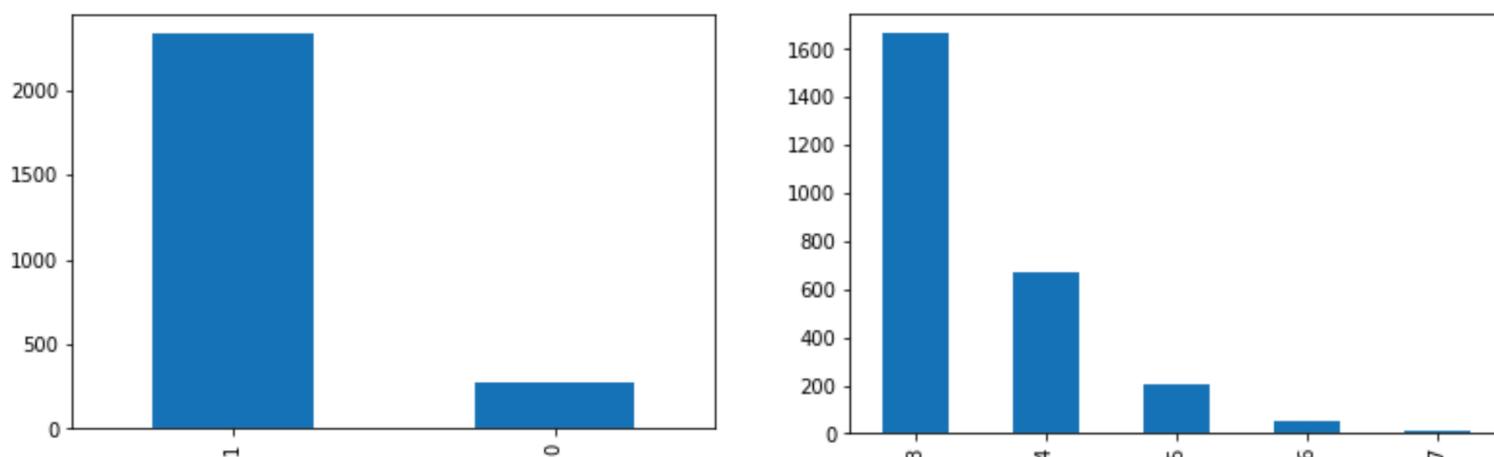
Conforme já alertamos em inúmeros momentos desta pesquisa, o corpus se encontra desbalanceado. O que pode ser visto na figura 25 (abaixo). Nela apresentamos uma representação gráfica da distribuição das categorias que foram

¹¹² Tradução nossa para: “neural networks are not only increasingly practical, they also promise non-trivial performance gains over traditional models. This is because unlike traditional models, which face performance plateaus even as more data become available, neural models continue to improve.”

¹¹³ É importante ressaltar desde já que, devido a inexistência da classe restabelecimento em nosso corpus e à hipertrofia do recrudescimento, bem como às questões levantadas por Tatit de que a atenuação tem características próprias, optou-se por classificar a atenuação como direção descendente.

utilizadas para classificar os textos. Combinamos dois gráficos presentes no apêndice para ocupar menos espaço e demonstrar o desbalanceamento em diferentes níveis.

Figura 25 - Representação gráfica das categorias tensivas no corpus de treinamento¹¹⁴



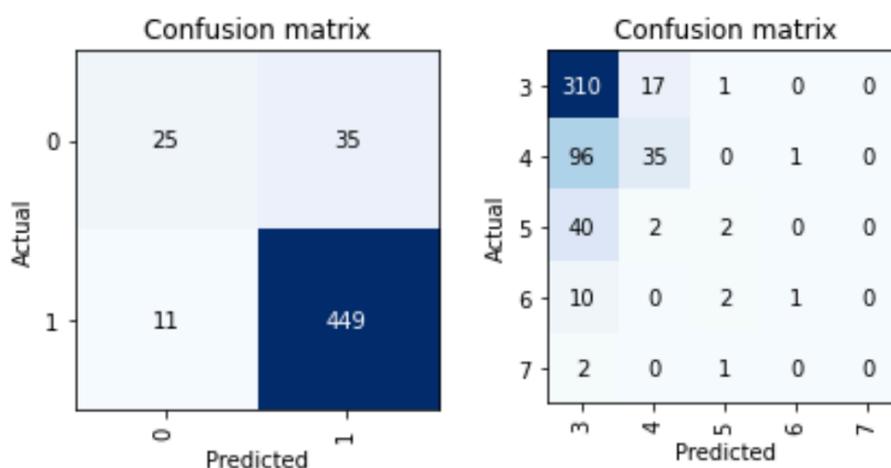
Fonte: produção própria a partir do Colaboratory e da linguagem Python de programação

Na parte esquerda da figura 25 temos a distribuição das avaliações de restaurante de acordo com a direção ascendente (1) ou descendente (0). No lado direito, as segmentações tensivas recrudescimento (3), saturação (4), atenuação (5), minimização (6) e extinção (7).

Os dados de treinamento, mais bem explicados no apêndice, dão conta de que o modelo obtêm alta acurácia devido ao desbalanceamento do corpus. Por conta da alta representatividade do recrudescimento, que oferece grande peso tanto na classificação por segmentos quanto por direções tensivas, o modelo apresenta baixa capacidade de generalização dos dados. O que pode ser visto na figura 26, abaixo:

¹¹⁴ Quantitativamente, temos para as direções: 2332 ocorrências ascendentes (1) e 270 descendentes (0); ao passo que para os segmentos: 1664 recrudescimentos (3); 668 saturações (4); 206 atenuações (5); 52 minimizações (6); e 12 extinções (7).

Figura 26 - Matrizes de confusão para direção (lado esquerdo) e segmento tensivo (direito)



Fonte: produção própria a partir do Colaboratory e da linguagem Python de programação

A leitura da figura faz-se da seguinte maneira: verticalmente (“Actual”), temos a classificação correta (proposta pelo analista), enquanto que horizontalmente (“Predicted”) temos a previsão da inteligência artificial. Quando o modelo está em ótimo funcionamento, a diagonal que parte da esquerda para a direita deve ficar o mais escura possível. De acordo com a figura 26, depreende-se que a inteligência tem mais de 90% de acurácia nas direções tensivas pois ela “chuta” mais ascendência (1) que é a classe abundante. Da mesma forma, em relação aos segmentos, pode-se ver que o canto superior esquerdo, no qual encontra-se a categoria recrudescimento, é o mais escuro. Por ser o mais populoso de forma exagerada, o modelo acaba querendo prevê-lo sempre.

Redes neurais performam mal em dados desbalanceados, um dos motivos pelo qual o modelo bayesiano ingênuo¹¹⁵ acaba sendo favorito como algoritmo de classificação. Entretanto, ele parece ter menor capacidade de dar conta dos novos ‘dialetos’ do português desenvolvidos nas redes sociais digitais. Além de classificarmos, manualmente, mais corpus para que o modelo funcione bem, novas pesquisas são necessárias para a demonstração de nossas hipóteses e a definição do melhor modelo de inteligência artificial para realizar nossa classificação semiótica.

¹¹⁵ Apresentado em 3.1.

4. CONCLUSÃO

O futuro da humanidade, segundo Hjelmslev na citação que apresentamos na introdução a este trabalho, poderia estar diretamente ligado à criação de uma ciência semântica que não desprezasse um lado prático a serviço dos seres humanos. Queremos crer, assim, que nossa pesquisa seja para o semioticista uma porta de (re)entrada no universo da ciência da computação e da inteligência artificial, rumo a realização daquela proposta.

As pesquisas estão apenas começando. A busca de competência para realizar os desafios maiores precisa ser contínua. Entretanto, optar pela análise do TripAdvisor parece ser¹¹⁶ uma boa estratégia introdutória para o neófito: a estrutura actancial dos textos é reduzida e gera uma classe de discursos bastante simples, com apenas uma categoria sendo manifestada.

Entretanto, os desafios maiores já estão batendo à porta: a invasão do Congresso Americano em 6 de janeiro de 2021 evidenciou que o relativismo linguístico é tamanho, que realidades paralelas (mundos possíveis?) já habitam entre nós. Centenas de cidadãos que arrombaram e depredaram o Capitólio dos Estados Unidos continuam a afirmar não entender o que fizeram de errado. Presos e investigados pela polícia, muitos afirmam que só estavam seguindo as ordens de Donald Trump.

(...) com os modernos meios postos à nossa disposição, o sistema de signos, a linguagem e a forma do conteúdo converteram-se numa força cuja aplicação nenhum governante poderá negligenciar. Hitler afirmou que poderia moldar a vontade das massas de modo a fazê-las mover-se como bem entendesse, e nem ele nem ninguém com semelhantes intenções poderia desconhecer a importância dos signos e símbolos para a realização de seus objetivos. O rádio, agora em combinação com a televisão, que amplia consideravelmente o efeito da palavra falada, é um importante instrumento da política internacional. Quem tiver o desejo e a habilidade de servir-se de tais aparelhos pode influir na vontade das massas... (HJELMSLEV; 1991: 107)

¹¹⁶ A escolha do objeto favoreceu nossos estudos. Outro tipo de texto, que já não tivesse todo esse aparato pronto (análise de sentimento, quantificações subjetivas etc), teria nos imposto duros desafios. O fato de o TripAdvisor ser composto majoritariamente de discursos do tipo sanção favorece muito a estreia na área das análises semióticas computadorizadas.

O vaticínio¹¹⁷ feito por Hjelmslev em 1953, como reitor da Universidade de Copenhague, não poderia estar mais certo - apesar de desatualizado. As redes sociais digitais e a democratização sem precedentes do /poder-fazer-saber/ acirraram ainda mais as afirmações do mestre dinamarquês. Após meses pregando que as eleições haviam sido fraudadas, reforçando cada vez mais a realidade paralela que parece habitar, Trump demonstrou uma das maiores contribuições de Saussure para a linguística: o signo é arbitrário.

A propaganda é capaz de explorar a língua como nunca se fez antes: é capaz de transformá-la, de adaptá-la às suas necessidades, e aquele que pretende tornar-se um ditador faria bem em estudar semântica. A linguagem está constituída de tal forma que novos signos podem ser formados constantemente a partir de componentes do signo, e como a relação entre forma e substância e entre conteúdo e expressão é arbitrária, a forma do conteúdo linguístico e o sistema semântico linguístico podem - desde que tenha suficiente habilidade e conhecimento - ser manipulados; e não só se podem criar novas palavras com novos sentidos, como ainda novos sentidos ou sentidos distorcidos podem ser imperceptivelmente incorporados a velhas palavras. (HJELMSLEV; 1991: 107)

Nosso trabalho, como dissemos, está apenas começando. Entretanto, assumir um lugar de destaque no processamento de linguagem natural nos parece ser para a semiótica uma vocação inata. Se Pierre Maranda queria saber quais figuras mais se manifestavam em uma tribo com a função de /amante adúltero/ e qual era o destino que os indígenas lhes davam, agora parece ser preciso descobrir como cidadãos discursivizam a democracia e a liberdade para antecipar golpes e desumanidades. A realidade não é uma existência indiferente ao homem, que a modela ao seu bel prazer.

4.1. Contribuições do trabalho

Acreditamos ter fornecido dados suficientes para encorajar novas pesquisas que busquem se aproveitar das interfaces entre semiótica e o processamento das grandes massas discursivas digitais. Apesar de não termos ido mais adiante em nossa pesquisa de mestrado, fornecendo dados quantitativos absolutos sobre a

¹¹⁷ “A forma de conteúdo da linguagem como um fator social”, in *Ensaio Linguísticos*; [1959] 1991; pp. 103-109.

plataforma TripAdvisor, acreditamos que esta possibilidade deva ser levada a cabo por pesquisas futuras. E mais, é preciso que novas plataformas sejam exploradas para que resultados possam ser comparados, da mesma forma como Maranda comparava os mitos de tribos diferentes.

4.2. Perspectivas futuras

Apesar dos problemas apontados em relação ao modelo de inteligência artificial proposto, já estamos realizando novas pesquisas junto à Escola Politécnica (Poli-USP) para otimização de nosso classificador. Esperamos, com isso, formalizar uma patente a ser dividida com a Universidade de São Paulo. Acreditamos que o desenvolvimento de patentes seja uma forma política de lutarmos pela autonomia universitária. Quanto mais formas de arrecadação encontrarmos para fortalecer nossas casas, mais tranquilas serão as noites da Ciência. E a intersecção entre semiótica e PLN nos parece promissora para o desenvolvimento econômico de inúmeras soluções computacionais.

Concomitantemente, estamos debruçados no estudo semiótico da rede social digital Twitter, muito mais complexa que o TripAdvisor (pois atualiza e realiza todo os componentes do esquema narrativo canônico). Já extraímos mais de 2 milhões de tweets que foram publicados durante a pandemia com os termos “covid”, “coronavírus”, “vírus” e “vacina”, pois acreditamos que seja uma boa porta de entrada para conhecer o fenômeno denunciado por Hjelmslev como distorção do sentido.

Junto às pesquisas realizadas por Lévi-Strauss, Cassirer, Barthes, Greimas e Zilberberg (que evidenciam acentos míticos conferidos aos discursos humanos), acreditamos ser necessário investigar os textos das redes sociais digitais à luz de “mitologias contemporâneas”. Não é difícil pensarmos em algo como “Levi.AI: uma inteligência semiótica de detecção dos graus de miticidade nos discursos da internet”. O sucesso de tal empreitada parece nos aproximar de Hjelmslev:

O conhecimento teórico puro sempre deve preceder o conhecimento aplicado. Porém, sem prejuízo de estudos mais profundos, deve o conhecimento estar sempre cômico do débito que tem para com o homem e a sociedade. (1991: 109)

REFERÊNCIAS BIBLIOGRÁFICA

ARNOLD, M.; **La machine molle de W. S. Burroughs: étude sémiotique assistée par ordinateur des structures narratives du roman**; Mémoire ou thèse; Université de Paris VIII; Paris; 1978.

_____ ; **La machine molle de W. S. Burroughs: étude sémiotique assistée par ordinateur des structures narratives du roman**; in *Semiotics Unfolding: Proceedings of The Second Congress of the Internacional Association for Semiotics Studies Vienna, 1979*; v. 1; pp. 713-720; De Gruyter Mouton; Berlin; 1984.

_____ ; **Report on the Semiotics and Artificial Intelligence Workshop** (Paris, France); in *The Semiotic Web 1987*; de Gruyter Mouton; pp. 667–700; Berlin; 1988.

_____ ; **La sémiotique: un instrument pour la représentation des connaissances en intelligence artificielle**; in *Études littéraires*; v. 21; n. 3; p. 81-90; Paris; 1989.

ASTI, P.; **Anotador morfossintático para o português-twitter**; Dissertação de Mestrado; PUC-Rio; Departamento de informática; Rio de Janeiro; 2011.

BACHELARD, G.; **A formação do espírito científico: contribuição para uma psicanálise do conhecimento**; Contraponto; Rio de Janeiro; [1938] 2016. Trad. Estela dos Santos Abreu.

BARROS, D.; **Teoria do discurso: fundamentos semióticos**; Humanitas; São Paulo; 2002.

_____ ; **A complexidade discursiva na internet**; in *CASA: Cadernos de Semiótica Aplicada*; v.13; n.2; p.13-31; Araraquara; 2015.

_____ ; **Uma reflexão semiótica sobre a "exterioridade" discursiva**; in *ALFA: Revista de Linguística*; v. 53; n. 2; [online]; 2009.

BEIVIDAS, W.; **A semiótica tensiva: uma teoria imanente ao afeto**; in *CASA: Cadernos de Semiótica Aplicada*; v. 13; n.1; p. 43-86; Araraquara; 2015.

_____ ; **Um modelo catenário e tensivo para a estrutura do quadrado semiótico**; in *Estudos Semióticos*; v. 15; 39-53; São Paulo; 2019.

_____ ; **Un modelo catenario y tensivo para la estructura del cuadrado semiótico: salir de Aristóteles**; in *Revista Tópicos del Seminario*, México, Puebla: [no prelo].

BENGFORT, B.; BILBRO, R.; e OJEDA, T.; **Applied Text Analysis with Python**; O'Reilly; EUA; 2018

FERREIRA, M.; e LOPES, M.; **Para conhecer linguística computacional**; Contexto; São Paulo; 2019.

FIORIN, J.; **Sendas e veredas da semiótica narrativa e discursiva**; in *DELTA*; v. 15; n. 1; São Paulo; 1999. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44501999000100009; Acesso: 22/01/2021

_____ ; **A festa do discurso: discurso e contexto sócio-histórico**. In *Entrepalavras*; v. 10; n. esp.; p. 12-26; Fortaleza; 2020.

FLOCH, J-M.; **A contribuição da semiótica estrutural para o design de um hipermercado**; In *Galaxia*; n. 27; p. 21-47; São Paulo [1987]; 2014. Trad. Ana Claudia de Oliveira

GOODFELLOW, I.; BENGIO, Y.; e COURVILLE, A.; **Deep learning**; MIT Press, EUA, 2016; Disponível em: <http://www.deeplearningbook.org>; Acesso: 22/01/2021

GREIMAS, A.; **L' énonciation: une posture épistémologique**; in *Significação: Revista de Cultura Audiovisual*; v. 1; pp 9-25; São Paulo; 1974; Disponível em: <http://www.revistas.usp.br/significacao/article/view/9011>. Acesso: 22/01/2021

_____ ; **A Enunciação (uma postura epistemológica)**. Cópia PDF traduzida por Jean Cristtus Portela e Maria Lúcia V. P. Diniz; disponível em: <http://semiotica.fflch.usp.br/sites/semiotica.fflch.usp.br/files/u45/enunciacao-Greimas.pdf>. Acesso: 22/01/2021

_____ ; **Semiótica e ciências sociais**; Cultrix; São Paulo; [1976] 1981. Trad. Álvaro Lorencini e Sandra Nitrini.

_____ ; **Sobre o sentido II**; Edusp, São Paulo; [1980] 2014. Trad. Dilson Ferreira da Cruz.

_____ ; **Da imperfeição**; Hacker Editores; [1987] 2002. Trad. Ana Claudia de Oliveira.

_____ ; e COURTÉS, J.; **Dicionário de semiótica**; Contexto; São Paulo; [1979] 2016. Trad. Alceu Dias Lima et al.

GOULARAS, D.; KAMIS, S.; **Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data**; in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*; pp. 12-17; Istanbul; 2019.

HJELMSLEV, L; **Ensaio linguísticos**; Perspectiva; São Paulo; [1959] 1991.Trad. Antônio de Pádua Danesi.

_____ ; **Prolegômenos a uma teoria da linguagem**; Perspectiva; São Paulo, [1943] 2018. Trad. J. Teixeira Coelho Netto.

HOCKETT, C.; **Review: Travaux du Cercle Linguistique de Copenhague Vol. V.”; in Recherches Structurales**; in *International Journal of American Linguistics (IJAL)* V. 18, N. 2, p. 86-99; Chicago; 1952

HOWARD, J.; GUGGER, S.; **Deep learning for coders with fastai & PyTorch - AI Applications Without a PhD**; O'Reilly; Canada; 2020

JURAFSKY, D.; MARTIN, J.; **Speech and language processing**; Prentice Hall; New Jersey; 2000.

KOTLER, P.; **Marketing 4.0**; Sextante; Rio de Janeiro; 2017. Trad. Ivo Korytowski

MALTAROLLO, V.; HONÓRIO, K.; SILVA, A.; **Applications of Artificial Neural Networks in Chemical Problems**; in *Artificial Neural Networks - Architectures and Applications*; pp 203-223; InTech; Croácia; 2013.

MARANDA, P.; **Formal analysis and inter-cultural studies**; in *Social Science Information*; v. 6; n. 4; pp. 7-36; Paris; 1967.

_____ ; **Semiotics and Computers: The Advent of Semiotronics?**; in *The Semiotic Web 1987*; de Gruyter Mouton; pp. 507–534; Berlin; 1988

_____ ; **Sémantographie du domaine "travail" dans la haute-ville et dans la basse-ville de Québec**; in *Anthropologica, New Series*; v. 20; n. 1/2; pp. 249-292; Canada; 1978.

MATTE, A.; LARA, G.; **Ensaio de semiótica: aprendendo com o texto**; Nova Fronteira; 2008 [sem numeração de página]

MOKARZEL,F.; SOMA, N.; **Introdução à ciência da computação**; Elsevier; Rio de Janeiro; 2008.

PANG, B.; e LEE, L.; **Opinion mining and sentiment analyses**; in *Foundations and Trends in Information Retrieval*; v. 2 n. 1-2; pp 1-135.; Boston; 2008 <http://dx.doi.org/10.1561/1500000011>

PONTI, P.; e COSTA, G.; **Como funciona o Deep Learning**; in *Tópicos em Gerenciamento de Dados e Informação*; Editora Sociedade Brasileira de Computação; pp.63-93; Uberlândia; 2017

SOUZA, P.; **O sujeito semiótico. Uma tipologia**; Tese de doutorado; Universidade de São Paulo; 2016.

TATIT, L.; **Quantificações subjetivas: Crônicas e críticas**; in *Caderno de Letras da UFF*; v. 1; n. 42; pp. 35-50; Niterói; 2011.

_____ ; **Passos da semiótica tensiva**; Ateliê Editorial; São Paulo; 2019.

_____ ; **Claude Zilberberg e a prosodização da semiótica**; in *Actes Sémiotiques*; [En ligne]. 2020, n. 123; Disponível em: <https://www.unilim.fr/actessemiologiques/6466>. Acesso: 23/01/2020

ZILBERBERG, C.; **Elementos de semiótica tensiva**; Ateliê Editorial, São Paulo; [2006] 2011. Trad. Ivã Carlos Lopes, Luiz Tatit e Waldir Bevidas.

_____ ; **Horizontes de la hipótesis tensiva**; Fondo Editorial/ UniLim; Peru; 2019.

_____ ; **La estrutura tensiva**; Fondo Editorial/ UniLim; Peru; 2016; Versão Kindle (ebook não paginado).

_____ ; **Razão e poética do sentido**; Edusp, São Paulo; [1988] 2006.

_____ ; e FONTANILLE, J.; **Tensão e significação**. Humanitas-USP, [1998] 2001. Trad. Ivã Carlos Lopes, Luiz Tatit e Waldir Bevidas.

A.POC: RELATÓRIO DE TREINAMENTO DE INTELIGÊNCIA ARTIFICIAL

Problemas de ordem técnica nos impossibilitaram, até o momento, de realizar de forma contínua a junção entre o relatório e este trabalho. Por conta disso, fomos obrigados a juntar nosso apêndice a partir da página seguinte. Devido a tais ocorrências, nossa prova de conceito não estará com as páginas numeradas.

PoC - Treinamento de Inteligência Artificial

O presente relatório faz as vezes de prova de conceito (ou PoC, na sigla do inglês, que significa Proof of Concept) para o modelo de inteligência que se pleitea. Ele foi gerado a partir da plataforma Google Colaboratory, que fornece maior capacidade de processamento ao mesmo tempo em que deixa nossas máquinas livres, não ocupando a memória durante a utilização.

Feito um gênero discursivo, relatórios como este (também chamados de notebooks) são cada vez mais comuns em apresentações de disciplinas ligadas as ciências de dados, processamento de linguagem natural, inteligência artificial entre outros. Por tanto, iremos nos valer da linguagem apresentada em inúmeros notebooks: apresentar resultados ao mesmo tempo em que se documentam os códigos de programação utilizados.

Sendo assim, partiremos da importação das bibliotecas, depois faremos uma breve análise descritiva dos dados e, por fim, do treinamento de uma rede neural para a classificação automática de avaliações no TripAdvisor - bem como de dados referentes à avaliação do modelo de deep learning.

O presente corpus faz parte de outro, muito maior, que conta com 45 mil avaliações de 31 restaurantes da cidade de São Paulo. Os dados, oriundos do TripAdvisor, recuam até o ano de 2016. Todos eles foram obtidos por meio de raspagem de dados (scrapers) realizada nos primeiros dias de fevereiro de 2020.

Desses 45 mil textos, classificamos manualmente apenas 2602, referentes a apenas dois restaurantes (Coco Bambu Anália Franca e Seen). A escolha dos restaurantes, apesar de aleatória, visou buscar a representação de classes diferentes. Uma vez que o primeiro deles, localizado na zona leste de São Paulo, está na faixa de preço de dois cifrões (preços médios); enquanto que o segundo, localizado em uma cobertura de hotel na Avenida Paulista, na faixa dos quatro (preços bastante elevados).

Importação de bibliotecas

Como primeira etapa, preparamos o ambiente de trabalho fazendo o download do pacote de Inteligência Artificial do FastAi

```
In [1]: !pip install -Uqq fastai
```

	194kB 5.7MB/s
	61kB 10.4MB/s

A seguir, realizamos as importações necessárias para exibir os dados que serão utilizados em nossa prova de conceito.

```
In [2]: from fastai.text.all import *
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Para sermos capazes de operar com dados externos (corpus, datasets etc) é preciso habilitar o acesso ao Google Drive. Por meio dos comandos abaixo, as operações junto ao drive serão autorizadas.

```
In [3]: from google.colab import drive
drive.mount('/content/drive')
```

Mounted at /content/drive

Acessamos o Google Drive para obter acesso ao arquivo QS(Poc)2.csv, que é a planilha com nosso corpus anotado.

```
In [4]: df = pd.read_csv("/content/drive/My Drive/QS(PoC)2.csv", delimiter=';')
```

Análise descritiva

Após (1) a configuração da plataforma, (2) a importação das bibliotecas que serão utilizadas, (3) o acesso ao Google Drive e a (4) obtenção do corpus, podemos realizar nossas primeiras análises. Nesta etapa, iremos nos valer apenas de algumas exibições, para apresentar nosso corpus, além de alguns gráficos, para fornecer estatísticas de nossa anotação.

Abaixo, é possível visualizar as cinco primeiras linhas de nosso dataset:

```
In [5]: df.head()
```

Out[5]:

		review_body	segmento	direção
0		almoço com amigos comemoração de final de ano, fomos bem acomodados, atendimento a contento, bebidas de acordo e comida boa e bem servida.	3	1
1		Jantar de negócios com serviço de primeira dos garçons :João Paulo e Wagner e o Somellier Wellington	3	1
2		Leonardo atendimento muito bom gente finíssima! Comida muito boa! Muito atenciosos! Parabéns! ¡!!!!	3	1
3		.minha primeira visita ao restaurante fui mt bem atendida e no final surpreendida com um pedido de namoro	3	1
4		1.000 excelente atendimento do gerente Charles juntamente c sua equipe...parabéns pára todos garçom souza nos auxiliou muito bem.	3	1

Nesta apresentação das cinco primeiras linhas de nosso corpus, é possível visualizar um exemplo da classificação. A notação em 'segmento' igual a 3 representa recrudescimento, enquanto que a direção 1 é a ascendência.

Abaixo, é possível visualizar as cinco últimas linhas de nosso dataset. Nelas, o único segmento presente é o da extinção (7) e a direção descendente (0). (Tanto acima, quanto aqui, fica claro que nosso corpus está organizado em ordem crescente dos segmentos.)

```
In [6]: df.tail()
```

```
Out[6]:
```

	review_body	segmento	direção
2597	Fomos hoje jantar 11:51 após consultar o site que o restaurante fecha a 1 hr.Chegamos ninguém nos atendeu então abordamos uma moça que suposto que era à hostess do local se fez de desentendida até por um momento achei que era uma cliente por tanto descaso ... depois de conversar disse que o bar era até 1h e o restaurante já estava fechado fez a menor questão de nos receber... o local é bonito mas tbm nada de mais para quem já conhece outros rooftop fora do Brasil.Enfim, não jantamos e também não voltarei pelo péssimo atendimento da recepcionista.	7	0
2598	Fui com minha esposa para jantarmos, ao chegarmos às 20:37 informaram que éramos o primeiro da fila, porém após abrir o aplicativo descobrimos que éramos o quarto da fila e com previsão de espera de 45 minutos e após 45 minutos fomos para 3 da fila e após mais 30 minutos continuávamos em 3 da fila e desistimos, neste tempo conversei com a recepcionista sobre o caso e a mesma não deu a mínima para a situação, não recomendamos pelo péssimo atendimento e descaso com o cliente	7	0
2599	Fui restaurante e fiquei estarecida com o péssimo atendimento da recepcionista Carol. Mal educada, mal humorada, foi grossa não só comigo, mas também com outras pessoas que perguntavam sobre a espera e o local.Infelizmente não tive nem o prazer de aguardar, pois não saio de casa para me deparar com tamanha falta de respeito de uma prestadora de serviço.	7	0
2600	Quando chegamos pedimos mesa para 04 e conforme solicitado pela hosttes descemos até o bar do hotel pois não havia lugar no bar na cobertura.Uma de nossas amigas chegou depois e foi direto ao bar superior imaginando que estávamos lá, sentou-se confortavelmente pediu um drink e aí fomos encontrá-la.Porém não tínhamos comanda, e quando o garçom pediu a HOSTTES, nossa comanda ela nos expulsou, falando que só poderíamos subir para o restaurante quando ela mandasse a mensagem autorizando, mas incrivelmente o bar tinha diversos lugares e no bar do hotel não tinha ninguém para subir!Eu nem ia pos...	7	0
2601	Chegamos no local por volta de 11:30h apenas para tomar drinks. A vista do local é realmente muito bonita. Após quase uma hora de espera sentamos no bar e ficamos aguardando atendimento. No bar tinham apenas dois funcionários para fazerem os drinks. Pois bem, após 20 minutos sentados a garçonete nos entregou o cardápio. Ficamos por mais de 30 minutos esperando que alguém nos atendesse e nada. Após quase 1h no local não fomos atendido e resolvemos ir embora. Esse foi o resumo da nossa noite no local. Espero que possam melhorar a atenção com os clientes.	7	0

Por meio do comando '.info()' é possível observar que o corpus tem 2602 linhas (são as amostras de avaliações de restaurantes), além de ser formado por três colunas. (O cabeçalho, com a discriminação das colunas, não é computado como linha).

```
In [7]: df.info()
```

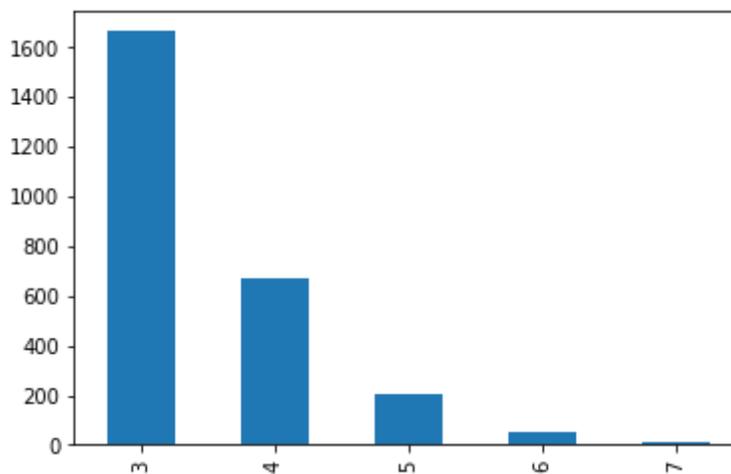
```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 2602 entries, 0 to 2601  
Data columns (total 3 columns):  
#   Column          Non-Null Count  Dtype  
---  ---  
0   review_body     2602 non-null   object  
1   segmento        2602 non-null   int64  
2   direção        2602 non-null   int64  
dtypes: int64(2), object(1)  
memory usage: 61.1+ KB
```

Na primeira coluna (review_body), encontram-se as avaliações de restaurante em formato de texto (ou object, uma vez que não limpamos o corpus, deixando-o igual foi produzido pelos usuários). Nas duas últimas temos, respectivamente, o segmento (recrudescimento, 3; saturação, 4; atenuação, 5; minimização, 6; e extinção,7) e a direção (ascendente, 1; ou descendente, 0). É possível perceber que ambas são compostas por dados em formato de números inteiros (int64), uma vez que optamos pela notação numérica para facilitar a classificação do corpus.

Abaixo, temos a distribuição dos **segmentos** de acordo com um gráfico de barras:

```
In [8]: df['segmento'].value_counts().plot(kind='bar')
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe16a5cd908>
```



Legenda (categoria x frequência):

```
In [9]: df.segmento.value_counts()
```

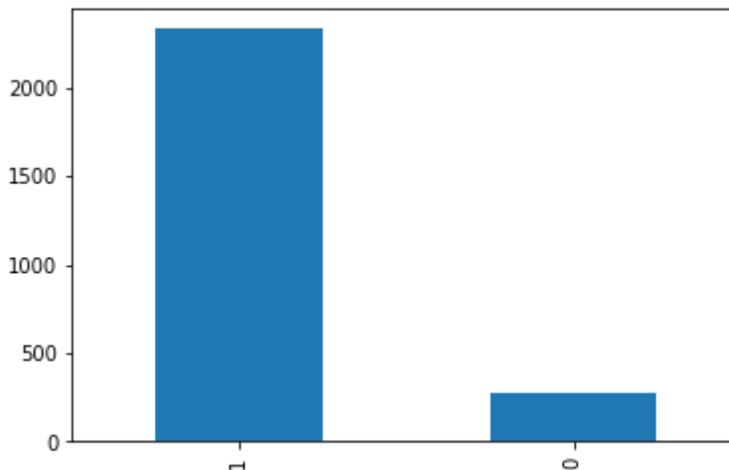
```
Out[9]: 3    1664
        4     668
        5     206
        6      52
        7      12
        Name: segmento, dtype: int64
```

Vemos claramente que o segmento 3 é, de longe, o mais presente em nosso corpus, representando 63,95%. O que já configura um desbalanceamento bastante prejudicial para o treinamento de uma inteligência artificial.

Abaixo, temos a distribuição das **direções** de acordo com um gráfico de barras:

```
In [10]: df['direção'].value_counts().plot(kind='bar')
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe16a49e908>
```



Legenda (categoria x frequência):

```
In [11]: df.direção.value_counts()
```

```
Out[11]: 1    2332
         0     270
         Name: direção, dtype: int64
```

Somados, os segmentos 3 e 4, únicos que formam a direção ascendente (1), representam 89,62% do corpus. O que comprova, mais um vez, o alto desbalanceamento das classes de nossos textos. O que implicará em overfitting, ou sobreajuste, que é quando a inteligência decora (mas não é capaz de aprender de verdade a realizar as classificações). Ou seja, apesar do alto índice de acerto nos dados de treinamento, tem performance ruim junto aos dados de avaliação, o que demonstra baixa capacidade de generalização.

Treinamento

Realizaremos dois tipos de treinamento utilizando o mesmo código para ambos. No primeiro, iremos colocar a coluna direção (0 ou 1) como Y, ou seja, como rótulo (alvo) da classificação que esperamos que a máquina seja capaz de reproduzir. Por fim, no último treinamento utilizamos os segmentos como alvo da classificação.

Y = Direção Tensiva

Primeiramente, precisamos configurar um objeto chamado 'dls' (conforme nomenclatura padronizada pelo FastAi) com alguns dados importantes. O código será definido a partir de alguns argumentos: nosso corpus (df), as avaliações como variável 'x' (text_col='review_body') e a 'direção' (label_col) como 'y' (que é o alvo da classificação). Aleatoriamente, por meio do argumento 'seed=42', iremos separar 20% de nossos corpus ('valid_pact=0.2'). Esse montante não será utilizado no treinamento, mas apenas no final, como dados novos, para verificarmos a capacidade de classificação de nossa inteligência artificial.

```
In [12]: dls = TextDataLoaders.from_df(df, text_col='review_body', label_col='direção', valid_pct=0.2, seed=42)
```

```
/usr/local/lib/python3.6/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
  return array(a, dtype, copy=False, order=order)
```

Abaixo é possível verificar uma amostra do que será processado:

```
In [13]: dls.show_batch(max_n=2)
```

		text	category
0		xxbos xxunk não gosta de um delicioso xxmaj brunch no domingo ? xxmaj aqui em xxup sp já fui em vários , no xxmaj terraço xxmaj xxunk (hotel xxmaj xxunk) , no xxmaj hilton e xxmaj xxunk xxmaj xxunk . xxmaj gostei de todos eles . xxmaj no xxmaj terraço xxmaj xxunk tem um buffet também com pães , frios , pratos quentes e sobremesas deliciosas que até parece um xxmaj brunch . xxmaj dessa vez resolvemos conhecer o xxmaj brunch do xxmaj restaurante xxmaj xxunk que fica no 23º andar do xxmaj hotel xxmaj tivoli xxunk saímos do elevador e seguimos para o restaurante , entramos num " xxunk " meio escuro de uns 2 xxunk de xxunk , tinha uma cortina e já xxunk uma xxunk , lembrando bem uma entrada de xxunk . xxmaj era xxunk , horário que começa o xxmaj brunch , e	0
1		xxbos xxmaj no sábado dia xxunk , fui com um grupo de 07 pessoas p / comemorar um aniversário . xxmaj xxunk às xxunk hs , já não havia lugar e nos recomendaram esperar no xxmaj must , bar xxunk , com bom atendimento . às xxunk hs , quando chegamos ao 23º andar , onde fica o xxmaj seen , fomos informados que a mesa só xxunk disponível após 01 hora de espera . a hostess , ao ser indagada sobre a xxunk de xxunk de uma das xxunk em xxunk da xxunk , se mostrou um pouco xxunk , mas disse que havia . xxmaj assim , de xxunk lugar na lista , xxunk para xxunk lugar na lista . xxmaj xxunk , ou não , fomos xxunk ao restaurante às xxunk horas , exatamente a tal 01 hora de espera . xxmaj enquanto os garçons do bar	0

Em nossa amostra de processamento é possível visualizar uma estranha notação sempre formada pela repetição da letra 'x'. Cada um daqueles termos representa uma característica oriunda da tokenização de palavras - redução para termos mais simples. 'xxbos'indica início de sentença; 'xxmaj', letra maiúscula da primeira letra; 'xxup', todas as letras maiúsculas; 'xxrep', repetição de caracteres (sempre acompanhado pelo número de vezes que se repete); enquanto que o 'xxunk' representa uma palavra incomum - que só apareceu uma vez no corpus.

Por não receber uma versão própria de vocabulário para tokenização, o código que utilizamos acaba inferindo tais informações a partir dele próprio. Portanto, quando percebemos que entre as palavras substituídas por 'xxunk' aparecem 'nublado', 'mara' (gíria que se refere ao adjetivo maravilhoso), 'sofás', 'juntos', 'enormes', entre outros, temos ainda mais evidências de que nosso corpus não está muito bom. Pois algumas palavras muito simples não estão conseguindo ser satisfatoriamente representadas.

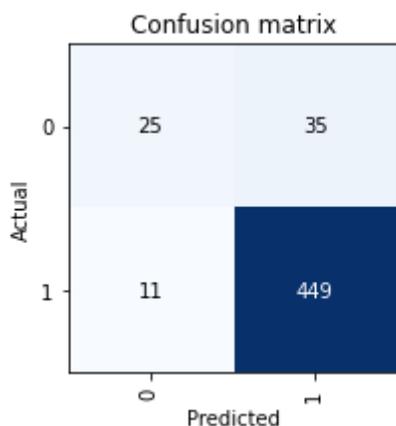
Ainda assim, realizamos nosso primeiro treinamento abaixo:

```
In [14]: learn = text_classifier_learner(dls, AWD_LSTM, metrics=accuracy)
learn.fine_tune(4, 1e-2)
```

epoch	train_loss	valid_loss	accuracy	time
0	0.690323	0.473411	0.880769	00:03
epoch	train_loss	valid_loss	accuracy	time
0	0.510278	0.339169	0.876923	00:04
1	0.388882	0.401258	0.867308	00:04
2	0.329815	0.277177	0.878846	00:04
3	0.290604	0.241814	0.911538	00:04

Nossa inteligência artificial foi composta por uma rede neural do tipo AWD-LSTM (ou ASGD Weight-Dropped LSTM) e utilizou a acurácia para avaliar a capacidade de classificação do corpus. Quatro épocas (epoch) de treinamento foram realizadas já partindo de uma taxa de aprendizado = 0.01 (ou 1e-2). Nos dados apresentados, podemos visualizar a rapidez com que o treinamento é realizado graças à disponibilização de GPUs pelo Google. Em nossa máquina, cada época estava levando em torno de 20 minutos - isso quando não travava e precisava ser reinicializada. Outro fator interessante é a alta acurácia, representada pelos 91% obtidos após a última época.

```
In [15]: interp = ClassificationInterpretation.from_learner(learn)
losses,idxs = interp.top_losses()
len(dls.valid_ds)==len(losses)==len(idxs)
interp.plot_confusion_matrix(figsize=(3,3))
```



Entretanto, quando solicitamos que uma matriz de confusão apresente os dados de nosso treinamento, vemos que as coisas não são tão boas assim. Afinal, nossa inteligência artificial até pode ser bastante boa na classificação da direção ascendente (1), acertando 97,6% (449 avaliações de 460), entretanto, quando precisa classificar a direção descendente (0) ela classifica corretamente apenas 41,6% (25 avaliações de 60) das amostras de tal tipo.

Observação: na matriz, é importante notar que 'Actual' representa a classificação correta, enquanto que 'Predicted' é a resposta dada pela máquina.

Y = Segmento Tensivo

Agora, iremos utilizar os 5 segmentos tensivos (recrudescimento, 3; saturação, 4; atenuação, 5; minimização, 6; e extinção, 7) presentes em nosso corpus como alvo de classificação (y ou label_col).

Em relação ao código, nada muda. Portanto, não iremos discutir ponto a ponto (como fizemos anteriormente), mas apenas apresentar um apanhado geral, após a matriz de confusão, e comentar a avaliação de nosso classificador.

```
In [16]: dls1 = TextDataLoaders.from_df(df, text_col='review_body', label_col='segmento', valid_pct=0.2, seed=42)
```

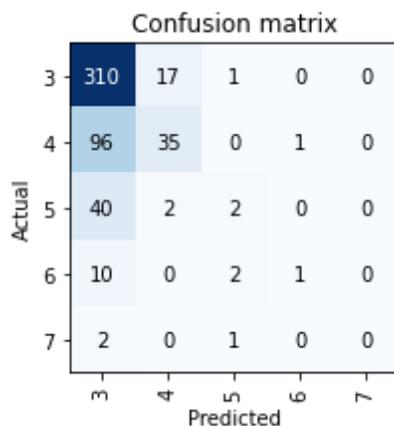
```
/usr/local/lib/python3.6/dist-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
  return array(a, dtype, copy=False, order=order)
```

```
In [17]: learn1 = text_classifier_learner(dls1, AWD_LSTM, metrics=accuracy)
learn1.fine_tune(4, 1e-2)
```

epoch	train_loss	valid_loss	accuracy	time
0	1.556350	1.135125	0.594231	00:03

epoch	train_loss	valid_loss	accuracy	time
0	1.163893	0.888965	0.644231	00:04
1	0.996555	0.891031	0.603846	00:04
2	0.887495	0.799327	0.675000	00:04
3	0.830536	0.787469	0.669231	00:04

```
In [18]: interp = ClassificationInterpretation.from_learner(learn1)
losses,idxs = interp.top_losses()
len(dls1.valid_ds)==len(losses)==len(idxs)
interp.plot_confusion_matrix(figsize=(3,3))
```



O desbalanceamento do corpus é ainda mais gritante quando tentamos configurar uma inteligência capaz de classificar cada um dos 5 segmentos tensivos. Em relação à extinção (7), por exemplo, o modelo não é capaz de acertar nenhuma predição (de 3). Quanto à minimização (6), prevê apenas 1 (de 13). Com relação à atenuação (5), a inteligência acerta apenas 2 de 44, o que representa 4,5% (os demais 95,5% são erradamente classificados como 91% recrudescimento e 4,5% saturação). Com relação aos segmentos pertencentes à direção ascendente, nosso modelo acerta 94,5% (310 de 328) dos recrudescimentos (3) e 26,5% (35 de 132) das saturações (4).

Grosso modo, o modelo parece apostar todas as fichas no recrudescimento: foram 458 palpites (predições) nesse segmento (3), 54 em saturação (4) e 6 em atenuação (5). Minimização (6) contou com 2 previsões (apenas 1 certa) e extinção não foi classificado. Ou seja, 88% das vezes ele "chutou" que seria recrudescimento (3), 10% saturação (4), 1,2% atenuação (5) e 0,8% minimização (6).

Com base nisso, acreditamos podemos dizer que fomos bem sucedidos em apresentar uma prova de conceito para o modelo de notação que estamos propondo. Entretanto, mesmo utilizando algumas das principais técnicas (no estado da arte) para a configuração de um modelo de deep learning, nosso modelo de inteligência artificial é um verdadeiro desastre.

Para reverter esse processo, acreditamos que o primeiro passo seja o balanceamento de nosso corpus. Para tanto, imaginamos que classificar manualmente apenas avaliações que tenham 1 ou 2 estrelas (deixando as de 5, 4 e 3 de fora) seja capaz de oferecer uma maior quantidade de amostras da direção tensiva descendente. Acreditamos, portanto, que um corpus balanceado seja a chave para que obtenhamos resultados mais satisfatórios para nossa proposta de inteligência artificial.