A FUNÇÃO DO ÍNDICE DE SÍNTESE DAS LINGUAGENS NA CLASSIFICAÇÃO GRAMATICAL COM REDES NEURAIS SEM PESO

Hugo Cesar de Castro Carneiro

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Rio de Janeiro
Agosto de 2012

A FUNÇÃO DO ÍNDICE DE SÍNTESE DAS LINGUAGENS NA CLASSIFICAÇÃO GRAMATICAL COM REDES NEURAIS SEM PESO

Hugo Cesar de Castro Carneiro

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Felipe Maia Galvão França, Ph.D.

Profa. Priscila Machado Vieira Lima, Ph.D.

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Thiago Alexandre Salgueiro Pardo, Ph.D.

Prof. João Carlos Pereira da Silva, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2012

Carneiro, Hugo Cesar de Castro

A Função do Índice de Síntese das Linguagens na Classificação Gramatical com Redes Neurais Sem Peso/Hugo Cesar de Castro Carneiro. – Rio de Janeiro: UFRJ/COPPE, 2012.

XII, 74 p.: il.; 29, 7cm.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2012.

Referências Bibliográficas: p. 56 – 63.

1. Redes neurais sem peso. 2. Classificação gramatical de palavras. 3. Arquiteturas neurossimbólicas. I. França, Felipe Maia Galvão *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

À minha esposa, Adriana.
Que estes sete anos juntos se transformem em mais sete. E estes em mais sete. E estes em mais outros sete ...

Agradecimentos

Inicialmente, e principalmente, agradeço à minha esposa Adriana, que me aguentou durante todo este tempo e que me ajudou a me focar e manter a calma quando eu achava que tudo estava perdido. Agradeço a ela também pelo auxílio e pelas noites em claro tentando arranjar os melhores termos para a composição do meu artigo e da minha dissertação. Seu companheirismo e sua cumplicidade sempre serão lembrados por mim!

À minha mãe pelo incentivo dado durante a fase inicial do meu mestrado e por acatar a minha opção pela pesquisa. Agradeço também por ela sempre procurar que eu me tornasse a melhor pessoa possível e por sempre querer estar presente em minhas conquistas.

Ao meu pai por todas as conversas durante minha adolescência. Elas me ajudaram a eu me tornar a pessoa que sou hoje. Uma pessoa correta, amiga e que tenta nunca se desviar de seus objetivos.

À minha vó, à minha tia e à minha irmã por todo apoio, reconhecimento e carinho dados.

Ao meu cunhado, André Mussap, que me auxiliou muito durante esta época conturbada da minha vida. A ajuda dele foi crucial para que esta fase fosse bem menos penosa.

A todos os meus familiares que presenciaram esta correria que foram estes últimos três anos e meio da minha vida.

Agradeço ao Prof. Felipe França e à Profa. Priscila Lima pela confiança que depositaram na minha capacidade, por terem apoiado a minha ideia de trabalhar com linguagens naturais e por todas as ideias prestadas que auxiliaram na confecção deste trabalho. Agradeço também a eles por continuarem crendo na minha capacidade e terem me aceitado como orientado de doutorado. Por fim, também os agradeço por terem me auxiliado nas vezes em que eu achava que não havia mais jeito, e nestas horas me mostrarem o caminho que eu deveria seguir.

Ao Prof. Thiago Pardo, ao Prof. João Carlos da Silva e ao Prof. Geraldo Xexéo, cujos comentários contribuiram valiosamente para a elaboração da versão final desta dissertação.

Meus agradecimentos à toda equipe do SIGA pelo companheirismo e pelo apoio

dado. Em especial para Gilson Tavares e Ricardo Storino, pela compreensão nas vezes que eu precisava me ausentar para participar de uma reunião ou ir a um congresso, e para Carlos Felippe Cardoso, que me auxiliou inúmeras vezes quando necessário. Agradeço muito mais que especialmente aos meus orientados Leopoldo Lusquino e Magno Ferreira que confiaram na minha capacidade de orientação e ao Prof. Felipe França que aceitou participar desta orientação comigo.

Meus sinceros agradecimentos às duas pessoas que informalmente me deram muitas ideias. Natália Giordani, que compreendia muito de linguística e estava, então, se aventurando no campo da computação. Ajudou com muitas ideias para minha dissertação de mestrado e também, indiretamente, para meu projeto final de graduação. E meu *e-pal* Clayton Cardoso que, tal como eu, é formado em computação, mas é fascinado pelo mundo da linguística. Me ajudou com muitos *insights* e me apresentou a vários *sites*, de onde pude pesquisar e obter muitas referências para meu trabalho.

Agradeço também aos colegas que fiz durante a minha graduação e o meu mestrado, dentre eles Douglas Cardoso, Cássia Novello, Gabriel Rosário, Prof. Luís Menasché, Bruno Lima, Daniel Alves, Marlon Rocha, Eduardo Ferreira, Fábio Couto, Horácio Lima, Antônio Benaion, Antônio Marcos Rojas, Raphael Simões, Tatiana Petra, Cláudio Miceli, Yanko Gitahy, Leandro Marques, Victor Castro, Paulo Braz, Glauber Marcius, Diana Rosa, Wagner Cavalcanti, Carlos Eduardo Grossi, Felipe Fraga, Sirius Thadeu, entre muitos outros.

Aos meus amigos de infância, de Jardim Sulacap, e os amigos que fiz no colégio e com quem ainda mantenho contato.

E, por fim, ao programa do PESC, à CAPES, ao CNPq e à FAPERJ pelo suporte dado a mim e a meus orientadores.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

A FUNÇÃO DO ÍNDICE DE SÍNTESE DAS LINGUAGENS NA CLASSIFICAÇÃO GRAMATICAL COM REDES NEURAIS SEM PESO

Hugo Cesar de Castro Carneiro

Agosto/2012

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Programa: Engenharia de Sistemas e Computação

A classificação gramatical de palavras em uma sentença é o ponto de partida para tarefas mais complexas, por exemplo, como a inferência de gramáticas. Para isso, as ferramentas responsáveis por tal procedimento precisam ter uma taxa de acerto muito elevada e um tempo muito baixo para classificar as palavras. As mais usadas hoje em dia utilizam abordagens que requerem o uso de processos iterativos, tornando a fase de treino muito lenta. Esta lentidão impossibilita a tarefa de classificação gramatical multilíngue, uma vez que, conforme o número de línguas cresce, o mesmo ocorre com o tempo de treinamento. Nesta dissertação se apresenta uma proposta para tornar o treinamento de classificadores gramaticais mais ágil. Primeiramente, propõe-se usar a WiSARD, uma arquitetura de rede neural sem peso, para executar as tarefas de classificação gramatical, uma vez que esta não necessita atingir qualquer convergência em seu treino. Ademais, cogita-se que haja uma relação direta entre os valores dos índices de síntese e os dos parâmetros da configuração ótima de rede para classificação gramatical em uma dada linguagem. Os experimentos conduzidos mostram que a arquitetura WiSARD realmente acelera a fase de treino dos classificadores gramaticais e também que se podem usar os valores dos índices de síntese para se estimar a melhor configuração da WiSARD capaz de executar classificações gramaticais em uma língua. Além disso, os experimentos também indicam que a classificação gramatical de palavras com a WiSARD é bem mais correta e precisa que a encontrada na literatura.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

THE ROLE OF THE INDEX OF SYNTHESIS OF THE LANGUAGES IN PART-OF-SPEECH TAGGING WITH WEIGHTLESS ARTIFICIAL NEURAL NETWORKS

Hugo Cesar de Castro Carneiro

August/2012

Advisors: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Department: Systems Engineering and Computer Science

Tagging parts of speech in a sentence is the foothold for more complex tasks, such as grammatical inference. In order to perform it, part-of-speech tagging toolkits need to have both a high accuracy value and a very low tagging time. Currently, the most commonly used toolkits make use of approaches that require time-consuming iterative processes, making the training phase too slow. This has hindered the adoption of multilingual part-of-speech tagging, since time spent in the training phase used to grow according to the number of languages. This dissertation presents a proposal to hasten the training phase of the part-of-speech tagging process. First, it is proposed to use WiSARD, a weightless artificial neural network architecture, to perform part-of-speech tagging tasks, as it does not need to achieve any convergence during the training phase. It is also considered that there is a direct relation between the index of synthesis values and those of the parameters of the optimal network configuration for part-of-speech tagging in a given language. The experiments conducted demonstrate that WiSARD architecture really hastens the training phase of part-of-speech taggers and that the index of synthesis values can be used to estimate the best WiSARD configuration capable of tagging parts of speech in any language. Furthermore, the experiments show that both the accuracy and the precision of part-of-speech tagging tasks with WiSARD are higher than the state of the art.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A tarefa de classificação gramatical é muito importante no ramo da linguística computacional e do processamento de linguagens naturais. A partir desta, muitas outras tarefas de maior complexidade podem ser realizadas. Análises sintáticas de sentenças, também conhecidas como *parsing*, por exemplo, requerem uma gramática de tamanho finito e, para isso, precisa-se de um sistema que transforme um conjunto expansível de palavras em um conjunto finito de símbolos[1]. Na tradução automática estatística também é necessário que se saiba as classes gramaticais das palavras, uma vez que isto pode ser fundamental na definição da ordem das palavras da sentença traduzida e em quais termos serão empregados na tradução, por exemplo, a palavra inglesa *like* que pode ser traduzida como *como* em português se sua classe for um advérbio ou uma conjunção, como *semelhante* (ou *parecido*) se for um adjetivo, ou ainda como *gostar* (ou alguma de suas conjugações no presente do indicativo, excetuando-se a da terceira pessoa do singular) se for um verbo.

Tal como visto acima, classificadores gramaticias (também chamados de etiquetadores gramaticais) são importantes para resolver problemas de ambiguidade gramatical e que seu resultado pode ser usado como entrada em outras tarefas (ou, ao menos, como um dado complementar). Devido a isso necessita-se que a exatidão de um classificador gramatical seja a maior possível, pois os erros provenientes uma classificação ruim podem se propagar rapidamente. Ademais, um classificador precisa ser ágil tanto em seu treinamento quanto em sua classificação. Os classificadores que se encontram na literatura possuem uma grande rapidez nas suas fases de classificação, mas esta agilidade falta na fase de treinamento. Isto faz com que a adoção de um modelo multilíngue torne-se mais difícil, pois conforme o número de línguas usadas aumenta, mais treinamento é necessário e, assim, a escolha

[1] Afirma-se aqui que o conjunto de palavras em uma língua é "expansível", pois a quantidade destas varia de tempos em tempos, criando-se gírias, neologismos ou novos termos para designar itens e/ou atividades que não existiam até então. Por exemplo, o vocabulário técnico na área de computação é composto por uma série de substantivos, adjetivos e verbos que só surgiram durante a segunda metade do século XX ou no século XXI.

de um conjunto suficientemnete grande de línguas torna-se improvável.

Tarefas de tradução automática requerem que se possa converter um texto de uma língua para outra mantendo seu sentido intacto. Inicialmente, pensou-se em criar uma arquitetura na qual de uma língua se chegasse a uma *interlingua* e desta pudessem fazer traduções para qualquer outra língua [2]. Contudo, este modelo veio a se tornar menos utilizado, devido ao surgimento do modelo estatístico, onde se traduz diretamente de uma língua para outra. O modelo estatístico, porém, requer uma quantidade de exemplos muito grande (esta quantidade crescendo quadraticamente conforme o número de línguas aumenta). Desta forma, sem um classificador gramatical que possua um tempo de aprendizado rápido, torna-se bastante complicado que se possa obter um sistema de tradução automática que abranja um conjunto suficientemente grande de línguas.

1.1 Objetivos e contribuições

Propõe-se neste trabalho a criação de um sistema de classificação gramatical que possa ser usado em uma vasta variedade de línguas e que possua um tempo de aprendizado curto. Este funcionaria como um ponto de partida para a criação de sistemas mais complexos que sejam comuns a várias línguas, tais como indutores de gramáticas, tradutores automatizados, entre outros. No entanto, devido às peculiares que cada língua possui, mostra-se improvável que se possa construir um classificador único para se usar em duas ou mais línguas. Todavia, é possível se criar um classificador que funcione para qualquer língua, desde que as peculiaridades desta sejam representadas em parâmetros deste classificador.

O classificador aqui proposto, nomeado mWANN-Tagger, executa seu processo de classificação baseando-se somente em uma janela de contexto (que contém as palavras adjacentes à que se deseja classificar) e em um "dicionário" com alguns termos principais (algumas palavras e algumas terminações). O comprimento desta janela e o tipo de termos que constarão no "dicionário" dependem da língua onde será feito o processo de classificação gramatical e, desta forma, da configuração de parâmetros do sistema.

No entanto, mesmo havendo tecnologias que tornam possível a construção de um classificador que seja bem ágil durante sua fase treinamento, a necessidade de se estimarem seus parâmetros pode fazer com que o processo de treinamento continue sendo lento. Por isso, este trabalho também propõe uma técnica $O(1)$ para se conseguir obter a melhor configuração de parâmetros do classificador. Para isto, optou-se pelo uso do índice de síntese da língua cujas palavras deseja-se classificar. O índice de síntese é uma medida proposta por GREENBERG [3]. Ele é usado para classificar as línguas de acordo com a quantidade média de morfemas que uma

Tabela 1.1: Tamanho do vocabulário em diferentes línguas

Tamanho do *corpus* (número de palavras)	Tamanho do vocabulário	
	Turco	Inglês
1 milhão	106547	33398
10 milhões	417775	97734

palavra qualquer deste idioma venha a possuir. Línguas com um índice baixo são chamadas de isolantes (ou analíticas), enquanto que as com índice alto são chamadas de sintéticas.

Para a escolha deste índice como medida que definiria a configuração de parâmetros ótima do classificador, partiu-se da hipótese que quão mais isolante for uma língua, isto é, quanto menor for seu índice de síntese, menos importantes serão as terminações das palavras e maior será o tamanho da janela de contexto, enquanto que para as línguas sintéticas o oposto tenda a ocorrer. Deduz-se isto do fato de nas línguas isolantes quase não existirem morfemas derivacionais nem inflexionais, fazendo com que a grande maioria das palavras destas línguas sejam compostas por uma simples raiz e, desta forma, venham possuir um grande número de homônimos. Por outro lado, as línguas sintéticas possuem uma grande gama de afixos para a criação de novas palavras, o que faz com que homônimos sejam raros e que o tamanho de seu vocabulário seja bem maior que o de línguas isolantes, tal como pode ser visto na Tabela 1.1 [4]. Devido ao grande vocabulário das línguas sintéticas e da alta possibilidade de se desejar classificar uma palavra não apresentada ao classificador na fase de treinamento, as terminações das palavras são importantes para a classificação gramatical nestas línguas, uma vez que estas normalmente representam afixos ou partes de afixos que carregam consigo alguma informação relevante sobre a classe gramatical da palavra a ser classificada.

Um exemplo que corrobora a hipótese supracitada é a classificação gramatical da palavra *like* para a língua inglesa (razoavelmente isolante) e de *gostamos* para a língua portuguesa (razoavelmente sintética). Na primeira há uma vasta possibilidade de classes, sendo necessária a inserção de um pronome adjacente a esta palavra para que esta seja classificada como verbo. A palavra *gostamos*, por outro lado, somente pode ser classificada como verbo, pois a terminação *-amos* já informa que esta é um verbo da primeira conjugação na primeira pessoa do plural.

Por fim, o uso do índice de síntese já tinha sido proposto anteriormente para tarefas de recuperação de informação [5]. Por outro lado, até então não se ouviu falar de tentativas de usá-lo em procedimentos de classificação gramatical.

1.2 Trabalhos relacionados

Segundo [6], cogita-se que o primeiro algoritmo para classificação gramatical tenha sido feito como uma parte do analisador gramatical do Projeto de Análise de Discurso e Transformações (TDAP – *Transformations and Discourse Analysis Project*) de Zellig Harris [7]. Desde então vários modelos foram criados para resolver os problemas de ambiguidade categorial. Os modelos mais usados recentemente são as gramáticas de restrições [8], os modelos estocásticos (por exemplo, os modelos ocultos de Markov [9–13], os modelos de Markov de máxima entropia [14, 15] e os campos aleatórios condicionais (CRF – *Conditional Random Fields*) [16–18]), árvores de decisão [19–22], modelos neurais [23], métodos de aprendizado baseado em transformações (TBL – *Transformation-Based Learning*) [24] e baseado em memória (MBL – *Memory-Based Learning*) [25], classificadores de máxima entropia [26–30], métodos baseados em máquinas de vetores de suporte (SVM – *Support Vector Machines*) [31], assim como variações do algoritmo de escolha de vizinho mais próximo [32] e do modelo *perceptron*, como o *averaged perception* [33, 34] e o *perceptron* bidirecional [35].

Por outro lado, a classificação gramatical multilíngue ainda é um tema pouco abordado. Dentre as publicações que utilizam este tema, destacam-se [36–38]. O sistema que é mencionado nestes artigos, contudo, depende bastante do número de línguas que são usadas, uma vez que este utiliza informações interlinguais (*cross-lingual*) para ajudar os seus classificadores a obter taxas de acerto altas. Para a obtenção destas informações interlinguais, este sistema requer que se use *corpora* paralelos em seu treinamento. Estas informações são usadas para a desambiguação categorial das palavras. Por exemplo, ao se obter uma ambiguidade ao se tentar classificar a palavra inglesa *like* seria necessário buscar as palavras equivalente nos *corpora* de outras línguas. Se, por exemplo, no *corpus* da língua portuguesa se encontrasse a palavra *semelhante*, a palavra *like* seria classificada como um adjetivo. Diferentemente deste classificador, o mWANN-Tagger utiliza o índice de síntese, que é uma medida da própria língua cujas palavras deseja-se classificar e, desta forma, não necessita de *corpora* paralelos para seu treinamento.

Classificadores gramaticais baseados em modelos neurais foram propostos desde [23]. Dentre estes, os mais comuns são os que utilizam modelos neurossimbólicos [39, 40]. Mais recentemente, um classificador gramatical baseado em redes sem peso foi proposto em [41]. No entanto, todos classificadores neurais criados até agora foram usados apenas para classificação gramatical monolíngue.

1.3 Estrutura da dissertação

Para melhor explicar como o mWANN-Tagger funciona, é preciso compreender os conhecimentos em que este se fundamenta. Com este propósito, sugere-se a leitura do Capítulo 2 para uma melhor compreensão da arquitetura WiSARD e de seu funcionamento, assim como do Capítulo 3, a fim de se entender melhor a medida de índice de síntese. Passados estes capítulos, começa-se a falar do classificador propriamente dito no Capítulo 4. Serão expostos no Capítulo 5 os experimentos usados para descobrir se o índice de síntese de uma língua é suficiente para que o processo de classificação gramatical de suas palavras seja correto, tal como a comparação do mWANN-Tagger com outras abordagens. A conclusão e a proposição de trabalhos futuros serão apresentados no Capítulo 6.

Capítulo 2

Redes neurais sem peso e modelo WiSARD

As redes neurais artificiais (RNAs) [42] constituem um modelo computacional baseado nas redes neuronais biológicas. Elas são utilizadas para executar tarefas que não possuam soluções determinísticas eficientes e que envolvam aprendizado e reconhecimento de padrões. Para tal, as RNAs simulam o comportamento biológico, utilizando ligações entre neurônios (nós) para transferência e processamento paralelo de informação.

As RNAs baseiam-se na ideia de que cada neurônio é capaz de fazer somente uma parte do processamento, sendo portanto necessária toda uma rede para que se possa executar uma tarefa mais complexa. O agrupamento dos neurônios na rede varia de arquitetura para arquitetura. Por exemplo, no modelo mais conhecido, o *perceptron* multicamadas [43], estes estão agrupados como um grafo direcionado separado em camadas, onde cada um destes neurônios está ligado a todos os presentes na camada seguinte, como mostra a Figura 2.1.

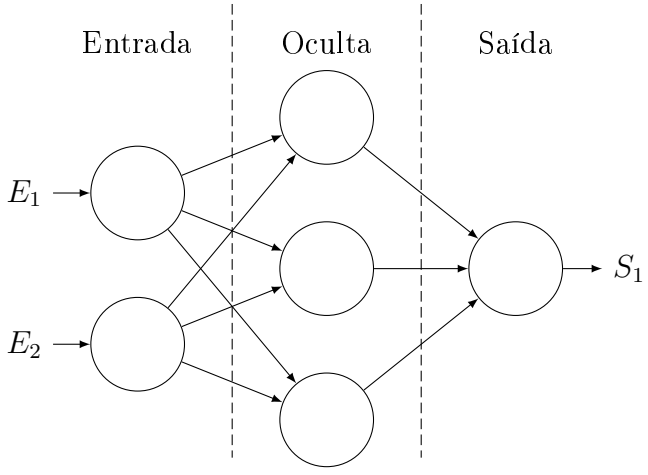


Figura 2.1: Rede neural *perceptron* multicamadas

2.1 Redes neurais tradicionais

As redes neurais tradicionais são caracterizadas por armazenarem quase todo seu conhecimento nas ligações entre os neurônios, denominadas sinapses. Os neurônios propriamente ditos somente contêm uma pequena parte do conhecimento, denominado tendenciosidade (ou "inclinação", denotado pela letra b – do inglês *bias*) [42].

As sinapses possuem pesos, os quais são responsáveis por modificar o valor de saída de um neurônio que será usado como entrada em outro. Estes pesos representam os tipos de sinapse: as inibitórias são representadas como pesos negativos, enquanto que as excitatórias como positivos. Durante a fase de treinamento, os pesos sinápticos são balanceados para que a rede consiga executar o procedimento desejado.

Para permitir um comportamento não linear para a rede, uma função de ativação monotônica não linear é atribuída à saída dos neurônios. As funções mais comumente aplicadas são a sigmoide e a tangente hiperbólica, devido ao seu comportamento assintótico e por sua derivada poder ser reescrita como função da mesma [42].

2.2 Redes neurais sem peso

Redes neurais artificiais sem peso (*Weightless Artificial Neural Networks* – WANNs) são um conjunto de modelos de redes neurais artificiais nas quais não há balanceamento de pesos sinápticos durante sua fase de treino. Esta ausência de peso nas sinapses é compensado pelo uso de memórias de acesso aleatório (*Random Access Memories* – RAMs) dentro de seus nós [44]. Tal como mencionado anteriormente, os neurônios das redes neurais tradicionais não guardam qualquer informação. Eles possuem somente uma variável de tendenciosidade e aplicam uma função não linear contínua. A troca destas funções pelo uso de RAMs faz com que as redes neurais sem peso sejam mais robustas e eficientes, uma vez que suas RAMs podem guardar qualquer valor, não requerendo continuidade, e que o treinamento dos dados nestas redes são simples operações de escrita em memória, desta forma não havendo convergência a ser atingida.

Existem vários modelos de redes neurais artificiais sem peso, por exemplo, WiSARD (***Wi****lkie* ***S****tonham and* ***A****leksander's* ***R****ecognition* ***D****evice*) [45]; AUTOWiSARD [46]; Memória Esparsa Distribuída [47]; *Probabilistic Logic Nodes* [48]; *Goal Seeking Neuron* [49]; G-RAM (*Generalizing RAM*) [50], sua implementação mais comum, a Virtual G-RAM (VG-RAM) [51]; *General Neural Unit* [52], entre outros.

2.2.1 O modelo WiSARD

A WiSARD é a rede neural booleana escolhida para ser usada como a base da arquitetura do classificador gramatical deste trabalho. Esta escolha deve-se à ausência da necessidade de se calibrar pesos (ou parâmetros), uma vez que se trabalharia com uma série de línguas, e nestas muitos testes seriam feitos. Ademais, este paradigma ainda é pouco avaliado, e em seu primeiro uso em tarefas de etiquetagem de classes gramaticais, em uma sondagem do seu uso na língua portuguesa, mostrou-se muito promissor [41].

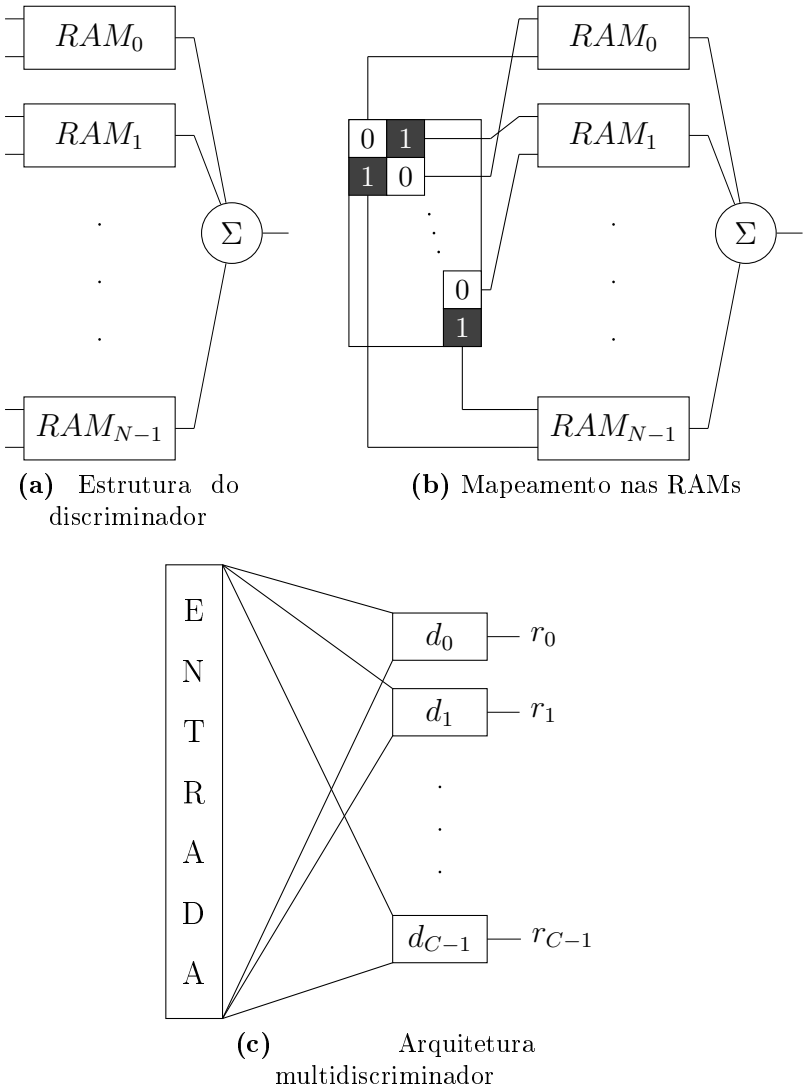
Como a rede WiSARD é um modelo booleano, esta somente tem a possibilidade de receber um conjunto de *bits* como entrada. Para uma mais fácil familiarização com a WiSARD, os exemplos usados neste capítulo envolverão o aprendizado de caracteres escritos, onde cada parte da imagem que contenha uma parte do caracter será representada pelo valor **1** e as restantes pelo valor **0**. A transformação das palavras a serem classificadas pelo etiquetador em *bits* será detalhada no capítulo 4, no qual é tratada toda arquitetura deste.

O elemento básico da WiSARD é o **discriminador-RAM** (ou somente discriminador), o qual é composto por um conjunto de RAMs e um somador Σ (ver Figura 2.2a). As RAMs guardam os dados treinados, inserindo **1** nas posições endereçadas durante a fase de treinamento e **0** nas outras, e o somador é responsável por informar a quantidade de RAMs cuja saída foi **1**. Este resultado é chamado de "medida de similaridade" da entrada em relação à classe do discriminador. Quanto maior for esta medida, mais similar a entrada é desta classe.

O conjunto de RAMs dos discriminadores é organizado de forma que cada *bit* da entrada seja associado a exatamente uma RAM através de um mapeamento pseudoaleatório (ver Figura 2.2b). Desta forma, se todas as N RAMs de um discriminador forem indexadas como RAM_i, o número de entradas n_i de uma RAM seguirá a relação

$$n_i = \begin{cases} \left\lceil \frac{M}{N} \right\rceil & \text{se } i < M \bmod N \\ \\ \left\lfloor \frac{M}{N} \right\rfloor & \text{se } i \geq M \bmod N \end{cases} \tag{2.1}$$

onde M é o número *bits* de entrada da rede, N o número RAMs do discriminador e i o índice da RAM em questão, que é um valor inteiro pertencente ao intervalo $[0, N-1]$. Pode-se verificar, a partir da relação 2.1, que a quantidade de entradas (e consequentemente o número de posições endereçáveis) das RAMs de um discriminador é inversamente proporcional ao número de RAMs que deste



(a) Estrutura do discriminador

(b) Mapeamento nas RAMs

(c) Arquitetura multidiscriminador

Figura 2.2: Arquitetura da WiSARD

discriminador venha a possuir.

A WiSARD, tal como qualquer rede neural, com ou sem peso, deve ser capaz de executar tarefas que envolvam reconhecimento de padrões. Para isto, ela precisa ter uma característica generalizadora que a permita classificar satisfatoriamente exemplos ainda não vistos. Esta característica pode ser garantida com o uso de diversas RAMs com poucas entradas, uma vez que a relação (2.1) garante que uma RAM tende a ser esparsa se seu discriminador possuir uma quantidade pequena delas, dado que essas terão um grande número de posições endereçáveis. Esta esparsidade faz com que as RAMs tendam a produzir uma saída **0** com muita facilidade caso a entrada da rede não seja extremamente parecida com alguma já aprendida por ela. Desta forma, aconselha-se o uso de discriminadores com várias RAMs, e estas com poucas entradas, em vez de somente algumas RAMs com muitas entradas.

O poder de generalização da rede WiSARD também é garantido pela existência do somador Σ. Este, ao informar o número de RAMs cuja saída foi **1**, pode gerar uma saída para o discriminador que varie de *0* até *N*, sendo *N* o número de RAMs que este discriminador possua. Desta forma, pode-se verificar que quanto menor o número de RAMs em um discriminador, maior será sua chance de produzir uma medida de similaridade baixa ao acaso. RAMs com muitas entradas contribuem com uma alta porcentagem da medida de similaridade. Por outro lado, as RAMs com poucas entradas apresentam um papel bem menos importante na determinação do valor desta medida, uma vez que seus discriminadores possuem muitas RAMs. Este fato corrobora a declaração de que a capacidade de generalização da rede é garantida com o uso de RAMs com pequeno número de entradas. Esta capacidade cresce de acordo com a quantidade de RAMs que um discriminador possua.

O discriminador, no entanto, só é capaz de informar se um padrão pertence a uma classe em particular ou quão similar ele é desta. A WiSARD, por outro lado, pode classificar um padrão em uma dentre diversas classes. Para fazê-lo, a WiSARD faz uso de uma **arquitetura multidiscriminador** (ver Figura 2.2c), na qual cada discriminador d_i informa a medida de similaridade da entrada à classe i. A classe escolhida é aquela cujo discriminador produzir a medida de maior valor, sendo, portanto, a mais similar ao padrão de entrada. Nas tarefas de classificação gramatical, cada classe representa uma classe gramatical distinta.

Fase de treinamento

A fase de treinamento da rede WiSARD consiste em duas partes, a de inicialização e o treinamento propriamente dito dos padrões de entrada. Durante a fase de inicialização atribui-se o valor **0** a cada posição de memória (ver Figura 2.3) e gera-se o mapeamento pseudoaleatório que associará cada *bit* da entrada (também chamada de "retina" em analogia ao reconhecimento de imagens) a uma RAM específica de cada discriminador.

Após isso, durante a fase de treinamento propriamente dita, obtêm-se os padrões de entrada a serem treinados (ver Figura 2.4a). Para cada padrão apresentado à rede, as posições das RAMs que forem acessadas durante este treinamento terão seu valor alterado para **1**, tal como mostra na Figura 2.4b. As posições somente são acessadas se pertencerem às RAMs do discriminador associado à classe sendo treinada, e se seu endereço for o que consta na tupla dos *bits* de entradas destas RAMs. Na Figura 2.4b os *bits* que compõem a entrada da RAM_0 formam a tupla **001**, a qual endereçam uma posição desta RAM, cujo valor é alterado para **1**.

A fase de treinamento termina após todos os padrões de entrada terem sido apresentados à rede, tal como é mostrado na Figura 2.4c. Diferentemente das redes com peso, nas quais a fase de treinamento requer que processos iterativos sejam

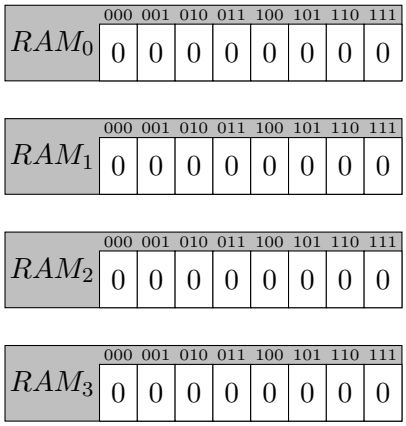
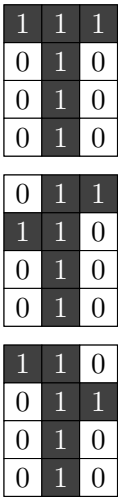
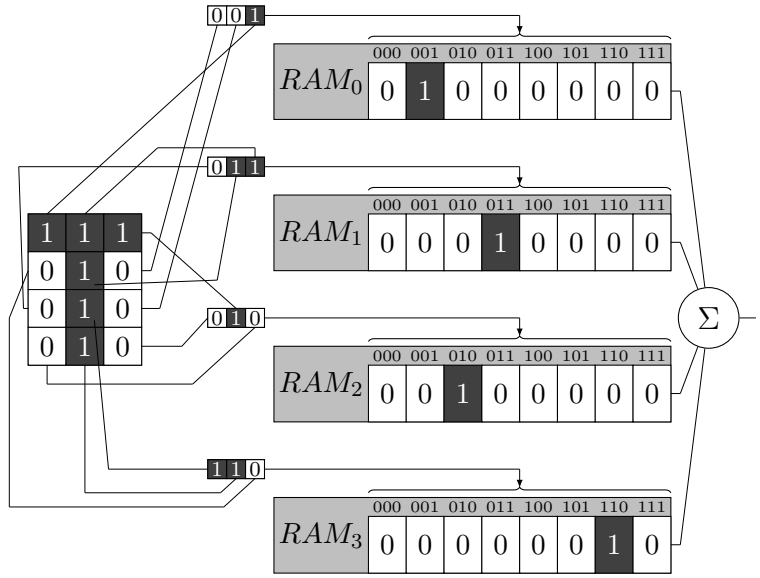


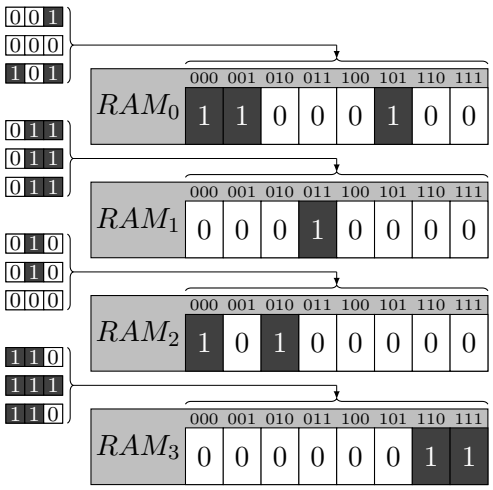
Figura 2.3: Configuração inicial das RAMs



(a) Exemplares de treinamento



(b) Treinamento de um exemplar



(c) Configuração das RAMs ao fim do treinamento

Figura 2.4: Treinamento da WiSARD

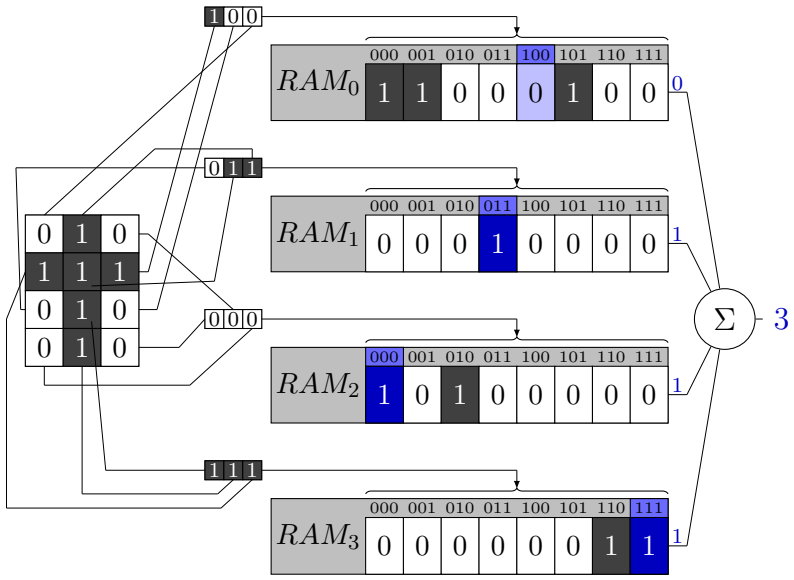


Figura 2.5: Classificação de um exemplar

feitos para que se atinja uma convergência, nas WANNs, por exemplo a WiSARD, somente é necessário que se apresentem os dados de treinamento uma vez para que estas já estejam devidamente treinadas.

Fase de classificação

Ao se apresentar à rede um exemplar ainda não visto por ela, os seus discriminadores tentarão classificá-lo, produzindo, então, uma medida de similaridade r_i. Esta medida é obtida contando a quantidade de RAMs cujas posições de memória foram endereçadas pelas k-tuplas associadas a elas. Na Figura 2.5, ao se tentar classificar um novo padrão de entrada, três das quatro RAMs de um dado discriminador retornaram o valor **1**, fazendo com que a medida de similaridade deste seja **3**.

Em uma arquitetura multidiscriminador, como mostrado na Figura 2.6, a classe que será associada ao exemplar submetido à classificação será aquela cujo discriminador produzir a medida de similaridade de maior valor r_{MAX}. A confiança γ desta resposta é calculada usando a fórmula $\gamma = \dfrac{r_{MAX} - r_{MAX-1}}{r_{MAX}}$, sendo r_{MAX-1} a medida com o segundo maior valor. Quão maior for o valor de γ, maior será a chance do padrão de entrada realmente pertencer à classe escolhida. O valor $r_{MAX} - r_{MAX-1}$, também denotado pela letra δ, representa a diferença entre as duas medidas de maior similaridade. Na Figura 2.6 pode-se perceber como funciona a análise de confiança de uma resposta. Nesta figura a maior medida de similaridade foi retornada pelo discriminador d_T e com segundo maior valor pelo discriminador d_I. A confiança da resposta corresponde à fração da resposta r_T que é superior a r_I.

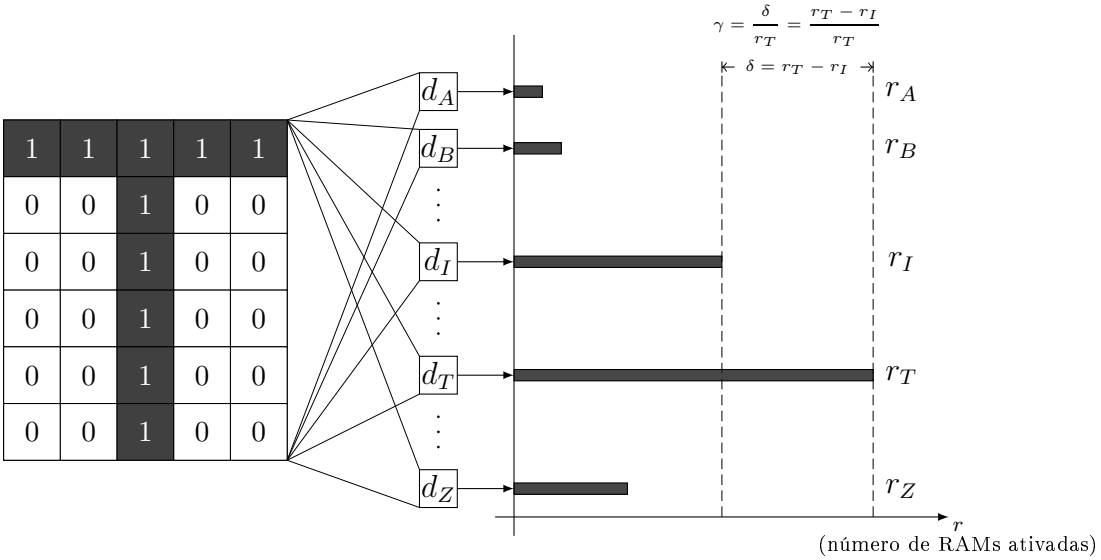


Figura 2.6: Classificação em uma arquitetura multidiscriminador

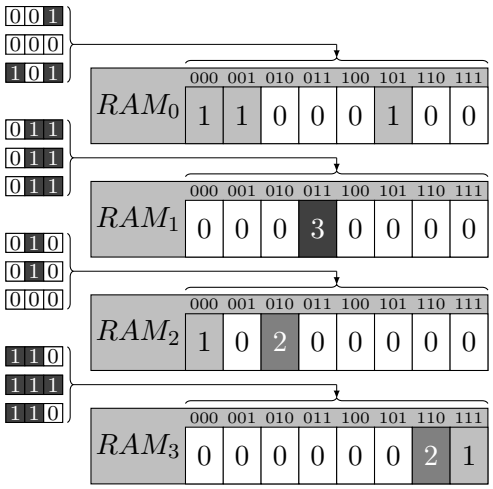


Figura 2.7: Treinamento de exemplares na WiSARD com a DRASiW

2.2.2 DRASiW: As "imagens mentais" e o fim dos empates

Devido à sua natureza de armazenamento de informação, os neurônios RAM permitem, através de um procedimento de engenharia reversa, a produção de exemplares, ou protótipos, derivados das classes que a rede aprendeu até então. Esta reversibilidade das WANNs foi primeiramente percebida em [53] e detalhada em [54]. A **DRASiW** é uma extensão da WiSARD que permite que um processo de engenharia reversa seja aplicado a uma rede deste modelo. Para que isso aconteça, as posições de memória têm de passar a guardar valores inteiros em vez dos tradicionais booleanos. Metaforicamente falando, em vez das posições de memória serem ou **pretas** (**1**) ou **brancas** (**0**), ao se usar a DRASiW estas passarão a ser **tons de cinza**, sendo preto o maior valor que possa aparecer em uma posição de memória de um discriminador e branco **0**.

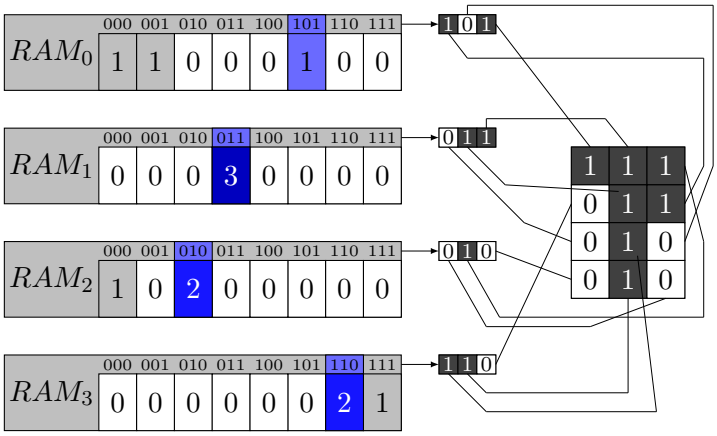


Figura 2.8: Geração de exemplar usando engenharia reversa

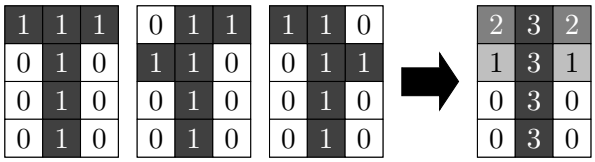


Figura 2.9: Geração de "imagens mentais" com DRASiW

Na Figura 2.7 é feita uma analogia com a configuração final das RAMs treinadas na Figura 2.4, representando como ficaria esta configuração final se a rede em questão utilizasse a DRASiW. Cabe notar que as posições que foram acessadas mais de uma vez passaram a possuir valores maiores que **1**. Para uma melhor compreensão, as posições menos acessadas estão em tons mais claros de cinza, enquanto que as mais acessadas estão em tons mais escuros.

Tal como mencionado anteriormente, a DRASiW permite a geração de exemplares a partir do conhecimento inserido dentro das RAMs. Para isto, necessita-se que se escolha a classe da qual se deseja gerar o exemplar e, então, para cada RAM do discriminador associado a esta classe, escolha-se uma de suas posições que já tenha sido acessada. Os endereços das posições acessadas são inseridos nas k-tuplas e, então, remapeados na "retina". Na Figura 2.8 pode-se perceber que este procedimento de engenharia reversa consta em basicamente executar um processo simétrico ao de treinamento. Neste exemplo escolheu-se a posição **101** na RAM_0, **011** na RAM_1, **010** na RAM_2 e **110** na RAM_3. A partir do mapeamento obteve-se o exemplar gerado.

Escolheu-se na Figura 2.8 as posições com maior número de acesso, pois em uma geração pseudoaleatória de um exemplar por este processo, estas posições serão as que terão maior chance de serem escolhidas. Isto faz com que os exemplares gerados a partir da DRASiW tenham uma maior probabilidade de serem próximos ao padrão mais comum. Fato que pode ser comprovado pela Figura 2.9, onde se tem o que seria uma "imagem mental" gerada pelos três exemplares de treinamento. Nota-se

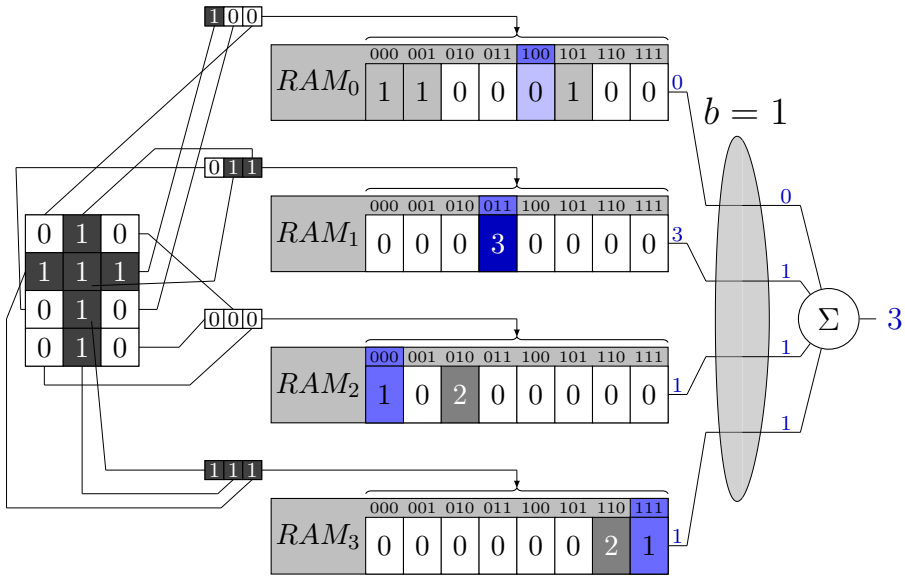
que as posições mais acessadas são as da coluna central, cada uma com 3 acessos, e as das duas extremidades superiores da "retina", com 2 acessos cada.

2.2.3 *Bleaching*

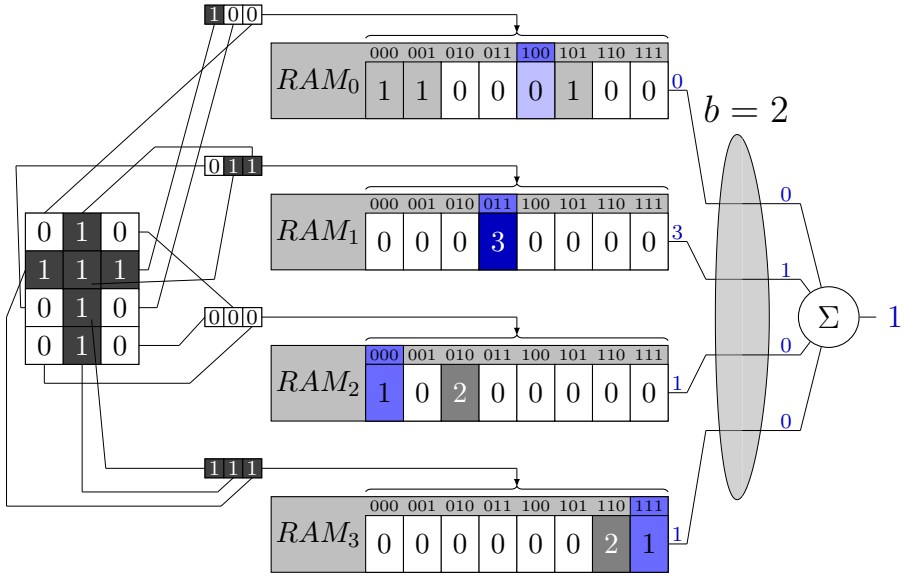
Em tarefas que envolvem reconhecimento de padrões, o aparecimento de dados ruidosos durante o aprendizado é bastante comum. Este tipo de dados, se aparecer com frequência durante o treinamento da rede, fatalmente fará com que seja atribuído o valor **1** a um grande número de posições de memória e, desta forma, tornando muito provável que dois ou mais discriminadores produzam a mesma resposta (alta). Para resolver tal problema, mostrou-se recentemente em [55] que a DRASiW pode ser útil no desempate das respostas dos discriminadores. Propôs-se a criação do ***bleaching***, uma técnica que usa em seu benefício a capacidade de produção de exemplares da DRASiW para os eventuais empates que apareçam nos procedimentos de classificação em redes treinadas exaustivamente.

No *bleaching* usa-se uma variável de limiar $b, b \geq 1$, a qual será atribuída à rede. Esta variável é usada para transformar, durante a fase de classificação, os valores inteiros retornados pelas posições de memória em **1**, caso estes sejam superiores ou iguais a b, ou **0**, caso contrário. Partindo-se de $b = 1$, este limiar é incrementado $(+1)$ enquanto ainda existirem empates nas respostas dos discriminadores. Uma vez que não haja mais empates, a classe associada ao exemplar recém-classificado será aquela cujo discriminador produzir a maior resposta.

Para uma melhor compreensão de como as funcionalidades da DRASiW e do *bleaching* se complementam, pode-se ver na Figura 2.10 que, após as RAMs retornarem valores inteiros, há uma região onde os valores maiores ou iguais a um dado valor são transformados em **1**, enquanto que os restantes são transformados em **0**. Comparando a Figura 2.10a com a Figura 2.10b nota-se que o acréscimo de 1 unidade no *bleaching* foi responsável em reduzir a resposta do discriminador em 2 pontos, pois tanto a RAM_2 quanto a RAM_3 passaram a possuir respostas inferiores ao limiar do *bleaching*.



(a) $b = 1$



(b) $b = 2$

Figura 2.10: Classificação na WiSARD com diferentes limiares de *bleaching*

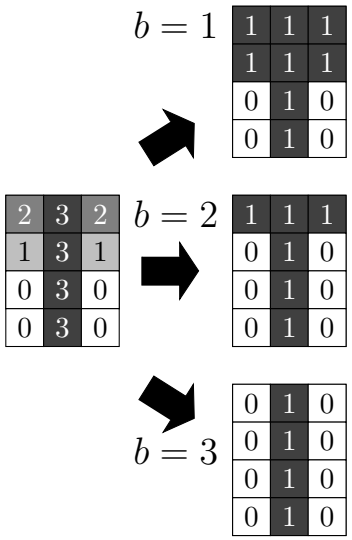


Figura 2.11: Uso do *bleaching* para transformar "imagens mentais" (em escala de cinza) em "preto e branco"

Metaforicamente falando, o *bleaching* é responsável por transformar as "imagens mentais" em tons do cinza da DRASiW novamente em imagens em "preto" e "branco". Tal analogia pode ser melhor notada na Figura 2.11, onde diferentes valores de limiar produzem diferentes imagens. A escolha do limiar é muito importante, pois limiares muito baixo podem fazer com que a imagem gerada esteja poluida com muita informação ($b = 1$) e os muito altos fazem com que informação útil se perca ($b = 3$).

2.2.4 *B-bleaching*

Uma WiSARD, no entanto, pode ser treinada com uma grande quantidade de dados, e nestes casos os empates persistem mesmo se b for um valor muito alto. O crescimento iterativo de b de uma unidade mostra-se como um procedimento ineficiente nestes casos. CARNEIRO *et al.* [41] propõem o uso de uma busca binária para determinar o melhor valor de b, $b \in [1, b_{MAX}]$, sendo b_{MAX} o maior valor presente em uma posição de memória. No entanto, devido a resultados obtidos em experimentos, nos quais os empates eram normalmente eliminados quando o limiar b era menor que $\sqrt{b_{MAX}}$, optou-se pelo uso da média geométrica em vez da média aritmética para determinar o funcionamento desta busca.

O Algoritmo 1 explica detalhadamente todo o processo de desempate com busca binária. Este algoritmo consta em achar um valor de b, no qual não haja empate entre os discriminadores e que a maior medida de similaridade encontrada em uma classificação seja igual à maior medida de similaridade obtida com $b = 1$. É importante ressaltar que, como a chance de empate é muito grande, no *b-bleaching* a diferença de 1 unidade entre as medidas de similaridade já é suficientemente

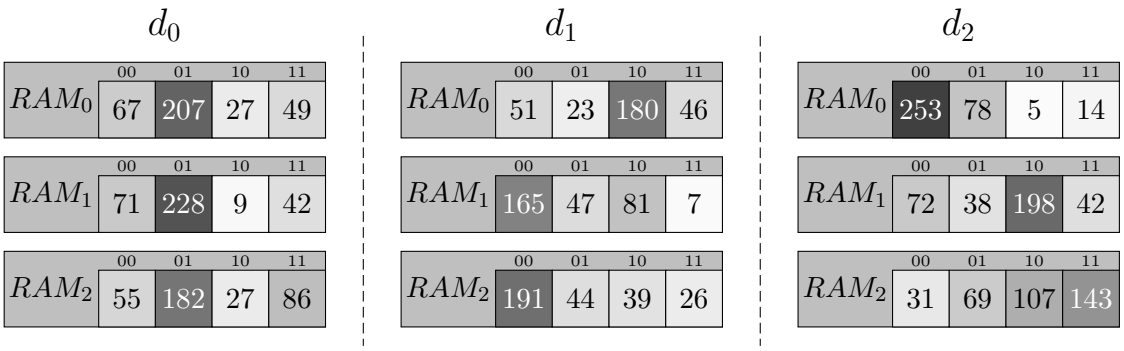
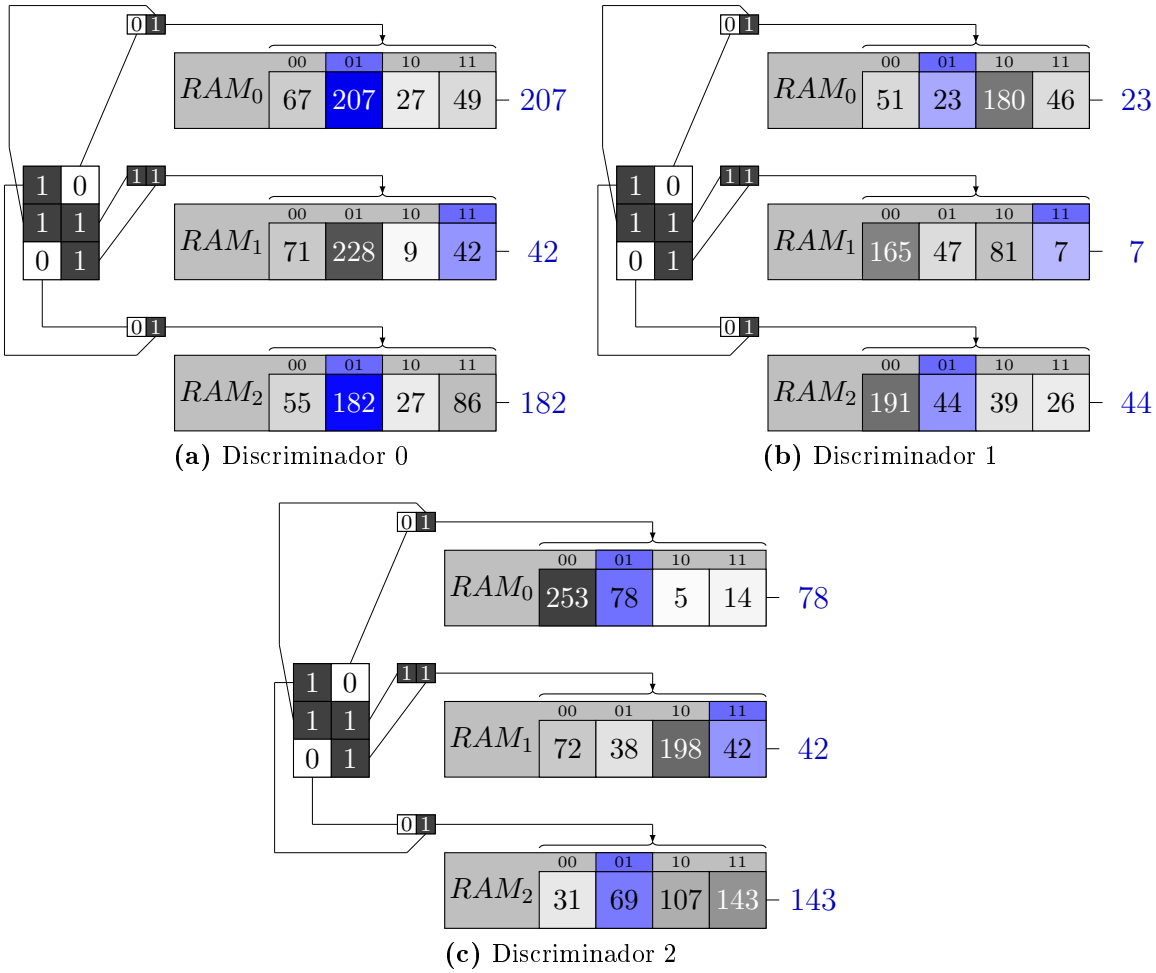


Figura 2.12: Configurações das RAMs no início da fase de classificação



(a) Discriminador 0

(b) Discriminador 1

(c) Discriminador 2

Figura 2.13: Obtenção do número de acessos nas posições acessadas das RAMs de cada discriminador

	b	r_0	r_1	r_2
	1	**3**	**3**	**3**
	253	0	0	0
$\lceil\sqrt{1 \times 253}\rceil =$	16	**3**	2	**3**
$\lceil\sqrt{16 \times 253}\rceil =$	64	**2**	0	**2**
$\lceil\sqrt{16 \times 64}\rceil =$	32	**3**	1	**3**
$\lceil\sqrt{32 \times 64}\rceil =$	46	**2**	0	**2**
$\lceil\sqrt{32 \times 46}\rceil =$	39	**3**	1	**3**
$\lceil\sqrt{39 \times 46}\rceil =$	43	**2**	1	**2**
$\lceil\sqrt{39 \times 43}\rceil =$	41	**3**	1	**3**
$\lceil\sqrt{41 \times 43}\rceil =$	42	**3**	1	**3**
$\lceil\sqrt{\mathbf{43 \times 253}}\rceil =$	**105**	**2**	0	0

Figura 2.14: Execução do *b-bleaching*

significativa.

As condições acima somente não conseguirão ser obtidas se existir um caso onde as duas maiores medidas de similaridade diminuem em uma unidade ao mesmo tempo. Para exemplificar tal possibilidade criou-se um exemplo onde deseja-se treinar exemplares compostos por 6 bits em uma rede com 3 diferentes discriminadores, cada um com 3 RAMs com 2 entradas cada. Considerando-se que a configuração inicial das RAMs ao fim do treino é a disposta na Figura 2.12 e que a Figura 2.13 mostra o processo de classificação de um exemplar, pode-se notar que os discriminadores d_0 e d_2 possuem a saída **42** como a menor obtida por alguma de suas RAMs (ver Figuras 2.13a e 2.13c) e que a menor resposta obtida por alguma das RAMs do discriminador d_1 é menor que este valor (a resposta obtida é **7**, pela RAM_1, como visto na Figura 2.13b). Devido às respostas retornadas pelas RAMs dos discriminadores, tem-se que não há um valor de b que faça com que alguma medida de similaridade sobressaia às outras e que seu valor seja máximo possível (neste exemplo, **3**), tal como ser visto na Figura 2.14.

Para resolver tal problema, assim que for averiguado que não há um valor de b, no qual não haja empate entre os discriminadores e que a maior medida de similaridade encontrada em uma classificação seja igual à maior medida de similaridade obtida com $b = 1$, esta última condição é alterada passa a ser que a maior medida de similaridade encontrada em uma classificação tenha de ser igual à maior medida de similaridade obtida com $b = 1$ menos 1 unidade, e assim por diante. Na última

linha da Figura 2.14 pode-se perceber que, ao se averiguar que para qualquer valor que b possua, não há como se conseguir uma medida de similaridade única que seja superior a todas as outras e que possua o valor $r_{ESP} = 3$, altera-se o valor esperado para esta medida de similaridade e recomeça-se a busca em um intervalo cujo limite inferior é o valor de b para o qual nenhum discriminador retorna uma medida de similaridade $r = 3$ (neste exemplo, $b = \mathbf{43}$) e o limite superior é igual ao maior número de acessos a uma posição de memória (neste exemplo, **253**).

Por fim, na Figura 2.15 pode-se ver o *bleaching* sendo usado nas saídas das RAMs da Figura 2.13 com diferentes valores de limiar que constam na Figura 2.14. Os extremos do espaço de busca, $b = \mathbf{1}$ e $b = \mathbf{253}$, podem ser vistos nas Figuras 2.15a e 2.15f, respectivamente. Na Figura 2.15b pode ser visto como o empate nos discriminadores d_0 e d_2 permanece após um passo do b-*bleaching*, quando $b = \mathbf{16}$. A queda simultânea nas medidas de similaridade tanto do discriminador d_0 quanto do d_2 pode ser vista nas Figuras 2.15c (quando $b = \mathbf{42}$) e reffig:bBleachingB43 (quando $b = \mathbf{43}$). O fim dos empates é obtido quando b obtém o valor **105** e a resposta do *bleaching* com este limiar é mostrada na Figura 2.15e.

Algoritmo 1: Executa desempate das respostas dos discriminadores com busca binária usando média geométrica

Entrada: b_{MAX}, o maior número de acessos a uma RAM
Saída: d_{MAX}, discriminador que emitir a maior resposta

Dados: d_{MAX-1}, discriminador que emitir a segunda maior resposta
Dados: r_{MAX}, maior resposta obtida por um discriminador
Dados: r_{MAX-1}, segunda maior resposta obtida por um discriminador
Dados: b, limiar do *bleaching* usado nos procedimentos de classificação
Dados: lim_{INF}, limite inferior da região de busca
Dados: lim_{SUP}, limite superior da região de busca

$lim_{INF} \leftarrow 1$
$b \leftarrow 1$
Executa classificação e obtém r_{MAX}, r_{MAX-1}, d_{MAX} e d_{MAX-1}
$r_{ESP} \leftarrow r_{MAX}$

se $r_{MAX} > r_{MAX-1}$ **então** `// Não houve empate com` $b = 1$
  retorna d_{MAX}
fim

$lim_{SUP} \leftarrow b_{MAX}$ `// Intervalo inicial de busca é` $[1, b_{MAX}]$

enquanto $lim_{SUP} > lim_{INF}$ **faça**
  $b \leftarrow \lceil \sqrt{lim_{INF} \times lim_{SUP}} \rceil$
  Executa classificação e obtém r_{MAX}, r_{MAX-1}, d_{MAX} e d_{MAX-1}
  se $r_{MAX} > r_{MAX-1}$ **então** `// Uma resposta sobressai-se às outras`
    se $r_{MAX} = r_{ESP}$ **então** `// ... e seu valor é o esperado`
      retorna d_{MAX}
    senão
      se $b = lim_{INF} + 1$ **então** `// Intervalo de busca pequeno`
        retorna d_{MAX}
      senão
        $lim_{SUP} \leftarrow b$ `// Reduz o intervalo de busca`
      fim
    fim
  senão
    `// Se nenhuma resposta sobressair`
    Executa atualizações nos valores de b, lim_{INF}, lim_{SUP} e r_{ESP} como consta no Algoritmo 2
  fim
fim

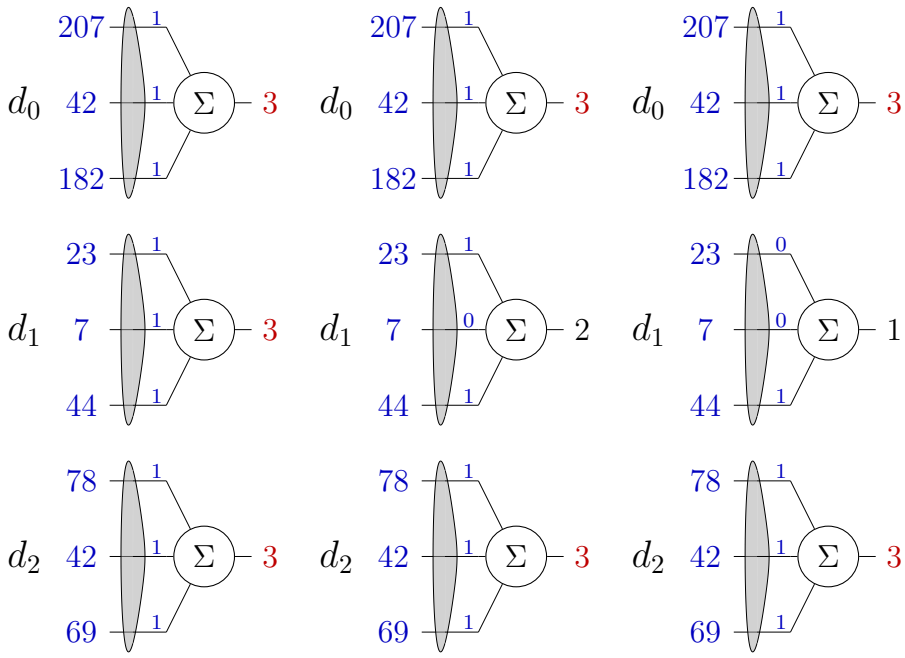
Algoritmo 2: Executa atualizações decorrentes da permanência de empate nas respostas dos discriminadores

Entrada: b_{MAX}, o maior número de acessos a uma RAM
Entrada: b, limiar do *bleaching* usado nos procedimentos de classificação
Entrada: r_{ESP}, resposta que se espera obter como a maior nos procedimentos de classificação
Entrada: r_{MAX}, maior resposta obtida por um discriminador
Entrada: lim_{INF}, limite inferior da região de busca
Entrada: lim_{SUP}, limite superior da região de busca

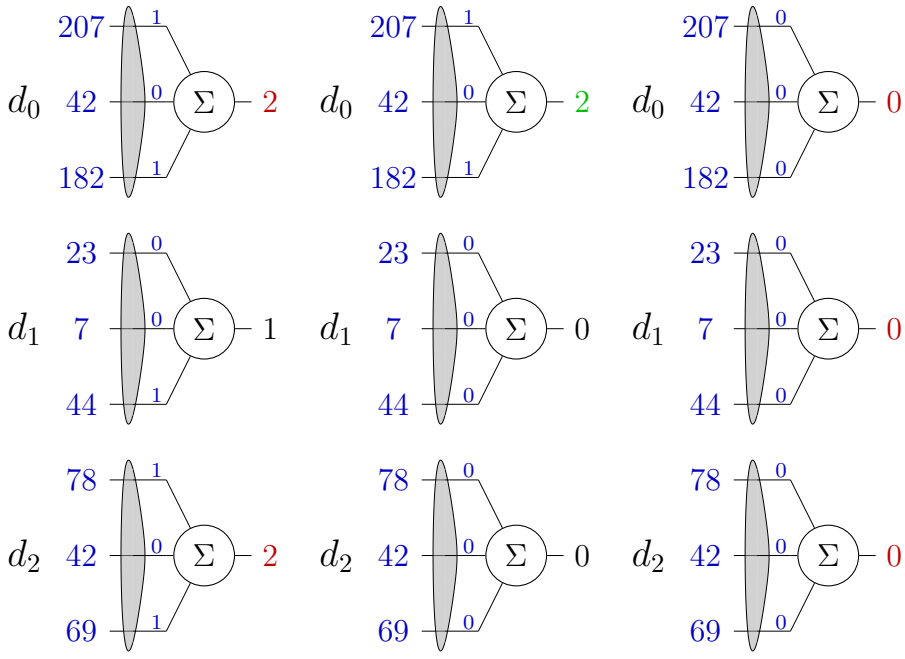
Saída: b, novo limiar do *bleaching* usado nos procedimentos de classificação
Saída: r_{ESP}, nova resposta que se espera obter como a maior nos procedimentos de classificação
Saída: lim_{INF}, novo limite inferior da região de busca
Saída: lim_{SUP}, novo limite superior da região de busca

```
// Se ainda há empates
```

se $r_{MAX} = r_{ESP}$ **então** `// A maior resposta é a esperada`
  se $b = lim_{SUP} - 1$ **então** `// Intervalo de busca muito pequeno`
    `// Atualiza o intervalo de busca para` $[b+1, b_{MAX}]$
    $lim_{INF} \leftarrow lim_{SUP}$
    $lim_{SUP} \leftarrow b_{MAX}$
    $b \leftarrow lim_{INF}$
    Executa classificação e obtém r_{MAX}, r_{MAX-1}, d_{MAX} e d_{MAX-1}
    $r_{ESP} \leftarrow r_{MAX}$
  senão
    $lim_{INF} \leftarrow b$ `// Reduz o intervalo de busca`
  fim
senão
  se $b = lim_{INF} + 1$ **então** `// Intervalo de busca muito pequeno`
    `// Atualiza o intervalo de busca para` $[b, b_{MAX}]$
    $lim_{INF} \leftarrow b$
    $lim_{SUP} \leftarrow b_{MAX}$
    `/* Diferentemente do bloco acima análogo a este, não há atualização de` b`. Este já se encontra com o valor para o qual seria atualizado no bloco acima. Pelo mesmo motivo também não há etapa de classificação. */`
    $r_{ESP} \leftarrow r_{MAX}$
  senão
    $lim_{SUP} \leftarrow b$ `// Reduz o intervalo de busca`
  fim
fim



(a) $b = 1$ **(b)** $b = 16$ **(c)** $b = 42$



(d) $b = 43$ **(e)** $b = 105$ **(f)** $b = 253$

Figura 2.15: Execução do *bleaching* em diferentes limiares

Capítulo 3

Índice de síntese

Segundo JURAFSKY e MARTIN [6], a morfologia é o estudo de como palavras são construídas a partir de unidades significativas menores, os morfemas. Estes morfemas são divididos em duas grandes classes: as raízes e os afixos, estes sendo ainda subdivididos em prefixos, os quais são adicionados "antes" da palavra (agrega-se à esquerda) que se deseja modificar, p.ex., o afixo *ante-* na palavra *antever*; sufixos, os quais são adicionados "depois" da palavra (agrega-se à direita), p.ex., o afixo *-eiro* em *padeiro*; infixos, que são raros na língua portuguesa e costumam representar vogais ou consoantes de ligação, p.ex., a consoante *-z-* em *cafezal*, e circunfixos, que não são usados na língua portuguesa, mas costumam ser usados em outras línguas, p.ex. o circunfixo *ge- -t*, que é usado para a formação do particípio passado de verbos regulares na língua alemã. O particípio passado do verbo *spielen*, por exemplo, é *gespielt*.

Existe uma variedade de formas de se combinar esses morfemas em palavras. Quatro destas são comuns e apresentam importantes papéis no processamento de fala e linguagem: a inflexão, a derivação, a composição e a cliticização. Esta, no entanto, será a única forma que não será tratada nesta dissertação, pois os clíticos são morfemas que agem sintaticamente como palavras, somente sendo reduzido em forma e anexado a outras palavras [6]. A preposição *d'* em *d'água* e o artigo italiano *l'* em *l'acqua* são exemplos de clíticos.

A quantidade de morfemas empregados para construir uma palavra e a forma como estes se combinam são dimensões que caracterizam morfologicamente as línguas. Para que se conseguisse ordenar e classificar as mais diversas línguas, propôs-se em [3] uma medida de topologia morfológica, denominada *índice de síntese*. Este índice, também conhecido como razão morfema-palavra, é uma escala numérica usada para ordenar as línguas de acordo com seu grau de síntese. As línguas podem ser classificadas de muito isolantes (ou analíticas), nas quais geralmente as palavras possuem somente um morfema, a polissintéticas, onde uma única palavra pode ter a semântica de uma sentença inteira em outras línguas. O

índice de síntese de uma língua pode ser obtido através da análise morfológica de um texto. Após a análise extrai-se seu valor usando a fórmula $IS = \frac{M}{W}$, sendo IS o índice de síntese, M o número de morfemas em um texto e W o número de palavras neste mesmo texto.

Este índice, por sua vez, pode ser subdividido em três índices menores (Na parte abaixo, W representa a mesma variável que na fórmula do índice de síntese):

- **Índice de síntese composicional** – Ele denota a preferência de uma língua pelo uso de termos derivados da composição de uma ou mais raízes, p.ex., *weekend* é uma composição, enquanto que seu equivalente em português, *fim de semana*, é composto por três palavras. Exemplos comuns de composição da língua portuguesa são vistos nas palavras *entrelinha*, *todavia* e *mandachuva*.

  O índice pode ser obtido usando a fórmula $CIS = \frac{R}{W}$, sendo CIS o índice de síntese composicional da língua e R o número de raízes em um texto.

- **Índice de síntese derivacional** – Mostra se uma língua faz um alto uso de afixos derivacionais para criar novas palavras a partir de alguma existente. Normalmente o processo de derivação, quando aplicado a uma palavra de uma determinada classe gramatical, cria uma nova com uma classe diferente da palavra original, p.ex., *feliz* é um adjetivo, mas *felic-idade* é um substantivo. Por outro lado, afixos derivacionais também podem alterar o significado de uma palavra, mantendo, entretanto, sua classe gramatical, p.ex., tanto *feliz* quanto *in-feliz* são adjetivos, mas o segundo termo é o exato oposto do primeiro.

  A obtenção deste índice faz-se usando a fórmula $DIS = \frac{D}{W}$, sendo DIS o índice de síntese derivacional da língua e D o número de afixos derivacionais em um texto.

- **Índice of síntese inflexional** – Este índice denota se uma língua tem predileção por exprimir diferentes categorias gramaticais, tais como número, caso, tempo, modo, pessoa, voz ou gênero, usando afixos (alto índice inflexional) ou clíticos e partículas (baixo índice inflexional). Diferentemente dos afixos derivacionais, os inflexionais não alteram a classe gramatical da palavra original, nem seu significado, p.ex., tanto *canto* quanto *cantei* representam o ato de *cantar*, mesmo que em tempos diferentes. A classe gramatical de ambos também é a mesma, verbo.

  Utiliza-se a fórmula $IIS = \frac{I}{W}$ para se obter este índice, sendo IIS o índice de síntese inflexional da língua e I o número de afixos inflexionais em um texto.

Abaixo são apresentados alguns exemplos de sentenças em línguas com diferentes graus de síntese. Cabe notar como o tamanho das palavras cresce, tal como sua capacidade expressiva, e como seus significados mudam do mais simples (às vezes atômico) nas línguas isolantes até muito complexos nas línguas sintéticas e polissintéticas:

- **Mandarim** (muito isolante)
  Míngtīan wǒ péngyou huì wèi wǒ zuò shēngri dàn'gāo.
  Decomposição da sentença: Amanhã eu amigo *futuro* para mim fazer aniversário bolo.
  Tradução: Amanhã meus amigos me farão um bolo de aniversário.

- **Inglês** (razoavelmente isolante)
  He travelled by hovercraft on the sea.
  Decomposição da sentença: Ele viajar-passado por *hovercraft* sobre o mar.
  Tradução: Ele viajou de *hovercraft* no mar.

- **Japonês** (razoavelmente sintética)
  Watashitachi ni totte, kono naku kodomo no shashin wa miseraregatai mono desu.
  Decomposição da sentença: Eu-plural em caso, esta chorar criança *possessivo* fotografia *tópico* mostrar-passivo-tender.a-negativo coisa ser.
  Tradução literal: No nosso caso, estas fotografias de crianças chorando são coisas que não tenderiam a ser mostradas.
  Tradução: Nós não conseguimos suportar sermos expostos a estas fotografias de crianças chorando.

- **Português** (consideravelmente sintética)
  Acreditaríamos neles se cumprissem suas promessas.
  Decomposição da sentença: Acreditar-condicional-1.plural em-ele-plural se cumprir-imperfeito.subjuntivo-3.plural 3.possessivo-feminino-plural promessa-plural.

- **Finlandês** (muito sintética)
  Käyttäytyessään tottelemattomasti oppilas saa jälki-istuntoa.
  Decomposição da sentença: Comportamento-tempo.não.passado-dele/dela obedecer-ausência-modo estudando-dever receber detenção-alguma.
  Tradução literal: O estudante cujo comportamento de agora em diante for de uma forma desobediente, receberá alguma detenção.
  Tradução: Quando um estudante se comportar de forma insubordinada, adquirirá uma punição.

- **Ainu clássico** (polissintética)
  Usaopuspe aeyaykotuymasiramsuypa.
  Decomposição da sentença: Vários-rumores 1.singular-aplicativo-reflexivo--aplicativo-longe-reflexivo-coração-balançar-iterativo.
  Tradução literal: Eu continuo balançando meu coração para longe e de volta para perto de mim sobre vários rumores.
  Tradução: Fico pensando a respeito de vários rumores.

Capítulo 4

mWANN-Tagger

O mWANN-Tagger (**M**ultilingual **W**eightless **A**rtificial **N**eural **N**etwork **Tagger**) é um etiquetador gramatical implementado na linguagem de programação Java, que pode ser usado para uma vasta variedade de línguas sem a necessidade de se implementar um processo para o tratamento da língua na qual deseja-se executar o procedimento de classificação gramatical. Ele surge como uma solução para o problema de desempenho que ocorre durante o treinamento dos classificadores. Isto se torna mais aparente conforme o número de línguas utilizadas cresce. Pretende-se usar tanto a robustez e eficiência das redes sem peso para tarefas de classificação quanto um método para se predizer de uma forma mais simples e direta os valores ótimos para os parâmetros da rede.

O mWANN-Tagger é uma evolução do WANN-Tagger [41], um etiquetador gramatical para a língua portuguesa, cuja arquitetura é baseada no modelo WiSARD. Ambos funcionam de forma similar, dado que ambos são etiquetadores gramaticais baseados em redes neurais booleanas e que usam tanto os últimos caracteres das palavras, de agora em diante denominados "terminações", quanto o contexto no qual a palavra a ser classificada está inserida. Entenda-se por contexto as A palavras situadas antes desta e as D depois. Tal como já mencionado na Seção 1.1, as terminações e o contexto são utilizados por estes classificadores para que, respectivamente, se possa classificar palavras não apresentadas no treinamento (comum na etiquetagem em línguas sintéticas) e para facilitar o processo de desambiguação em palavras que apresentam muitos homônimos (bastante comum em línguas isolantes).

Por outro lado, o mWANN-Tagger dispõe de alguns parâmetros a mais que o WANN-Tagger, todos dependentes da língua cujas palavras das sentenças pretende-se etiquetar. Os valores destes parâmetros são estimados de acordo com os índices de síntese da língua em questão.

Tabela 4.1: Conjunto de classes gramaticais do mWANN-Tagger

N	Substantivo	ADJ	Adjetivo
ADV	Advérbio	V	Verbo
PRON	Pronome	DET	Determinante
ADP	Adposição	NUM	Numeral Cardinal
CJ	Conjunção	MW	Classificador Numérico (do inglês **M**easure **W**ord)
PART	Partícula	INTJ	Interjeição
PUNC	Pontuação	MISC	Diversos

4.1 Conjunto de classes gramaticais

Para se criar um classificador gramatical multilíngue, mostra-se necessário o uso de um conjunto de classes gramaticais comum a todas as línguas usadas. Este conjunto surge como uma forma de manter um número fixo de classes, dada a necessidade de tanto se ter uma correlação entre as classes gramaticais das diversas línguas quanto haver uma normalização dos seus conjuntos de classes. Tal procedimento faz com que somente os parâmetros do classificador mantenham-se variáveis, uma vez que estes são os únicos valores que se deseja que sejam dependentes da língua.

Além disso, PETROV *et al.* [1] também expôs outras razões importantes para se usar um conjunto de classes gramaticais comum a várias línguas em vez de um específico para cada. Dentre elas destacam-se uma possível melhora nas comparações de resultados de etiquetadores para línguas diferentes e a possibilidade de serem criados etiquetadores para várias línguas com um conjunto de classes comum, facilitando o desenvolvimento de novas aplicações interlinguais, sem haver a necessidade de se manter regras ou sistemas específicos para uma ou outra língua devido a diferenças nas diretrizes de anotação das bases.

O conjunto de classes gramaticais usado para o mWANN-Tagger (Tabela 4.1) é similar ao usado em [1], exceto pela existência no conjunto proposto nesta dissertação de uma classe à parte para classificadores numéricos, já que esta é uma classe gramatical importante em línguas do extremo oriente, e de outra para interjeições. Um conjunto similar de classes gramaticais comuns também foi criado em [38].

Devido às diferenças entre as línguas, algumas etiquetas presentes no conjunto de classes da Tabela 4.1 podem não ser utilizadas por algumas línguas. Por exemplo, os classificadores numéricos são comuns nas línguas do extremo oriente, mas inexistentes em qualquer outra região do mundo. Esta classe representa uma "unidade" que precisa ser inserida entre um numeral cardinal e o item que se deseja contar, p. ex., em mandarim a sentença "*sān zhī māo*" significa "três gatos", onde *sān* significa **três**, *māo* **gato** e *zhī* é o classificador numérico para animais.

Tabela 4.2: Exemplos de formas adjetivas e substantivas dos pronomes

Tipo de pronome	Formas substantivas	Formas adjetivas
Possessivo	meu (*my*), deles (*their*)	meu (*mine*), deles (*theirs*)
Demonstrativo	este, aquela	isso, aquilo
Interrogativo	qual, quantos	quem, quê
Indefinido	algum, nenhum	alguém, nada

Outras etiquetas representam classes com funções análogas em diferentes línguas. Por exemplo, a etiqueta ADP, que representa a classe das adposições, abrange as preposições em algumas línguas e as posposições em outras. Estas possuem as mesmas funções em uma sentença, somente sendo diferenciadas pela posição onde estas são inseridas (as preposições à esquerda dos termos que elas modificam e as posposições à direita), p. ex., na língua turca a sentença "*senin için*" significa "para você", onde a preposição **para** é representada pela posposição *için*.

Destaca-se ainda a classe dos determinantes, cujos elementos são palavras que definem, especificam ou quantificam substantivos (ou sintagmas nominais) adjacentes a essas. Na gramática normativa esta classe é subdividida em *a*) artigos, que servem para determinar ou indeterminar substativos, p. ex., **a**, **o**, **um**, entre outros, e *b*) pronomes adjetivos, que especificam ou quantificam substantivos. Estes são subdivididos em pronomes adjetivos possessivos, demonstrativos, interrogativos e indefinidos.

Diferentemente dos pronomes adjetivos, os pronomes substantivos, que são marcados pela etiqueta PRON, substituem substantivos (ou sintagmas nominais). Este conjunto abrange os pronomes pessoais (**eu**, **ele**, **mim**, **se** etc) e os pronomes substantivos possessivos, demonstrativos, interrogativos e indefinidos. Estes últimos quatro tipos de pronome possuem diferentes formas adjetivas e substantivas. A Tabela 4.2 mostra exemplos de formas adjetivas e substantivas destes pronomes. Cabe notar que as formas adjetivas e substantivas dos pronomes possessivos são idênticas na língua portuguesa (**meu**, **vosso**, **deles**), mas são diferentes em outras línguas, tal como a inglesa (*my* e *mine*, *their* e *theirs*).

Por fim, as etiquetas PART e MISC, que representam as classes das partículas e de termos diversos, respectivamente, são usadas para etiquetar termos que não podem ser associados a nenhuma das classes restantes. A etiqueta PART é associada a palavras de classe fechada (i.e., de uma classe onde não seja possível o surgimento de novos termos) que não possam ser associadas a nenhuma das outras

classes fechadas (pronome, determinante, adposição, numeral cardinal[1], conjunção, classificador numérico e interjeição). A etiqueta MISC é associada a palavras em sentenças estrangeiras ou a termos presentes em um texto que não fazem parte da sentença onde este esteja inserido, por exemplo o advérbio latino *sic*, que é inserido em um texto para evidenciar um erro cometido pelo autor de uma sentença. Este termo, no entanto, não faz parte da sentença.

4.2 Treinamento

Para se treinar uma instância do mWANN-Tagger para uma determinada língua necessita-se que exista um *corpus*, no qual suas sentenças estejam anotadas na forma **[palavra|etiqueta]**. Por exemplo, a sentença "As razões para investir." deve aparecer no *corpus* como "[as|DET] [razões|N] [para|ADP] [investir|V] [.|PUNC]".

O mWANN-Tagger utiliza uma rede WiSARD como parte de sua arquitetura para fazer o procedimento de etiquetagem. Como esta é uma rede booleana, é necessário que as palavras que constam no *corpus* sejam transcodificadas em *bits*, para que estes possam ser utilizados como entradas da rede. Esta transcodificação ocorre em três passos durante a fase de pré-processamento dos dados.

O primeiro passo a ocorrer na fase de pré-processamento é a criação de um arquivo de mapeamento. Este é utilizado pelo etiquetador para transformar as palavras do *corpus* em um vetor com as probabilidades destas pertencerem a cada uma das C classes gramaticais (no conjunto em comum apresentado na Tabela 4.1, $C = 14$). Este arquivo contém, em cada uma de suas linhas, uma palavra encontrada no *corpus* e as probabilidades desta pertencer a cada uma das classes à sua direita. Abaixo destas estão dispostas as terminações consideradas como relevantes na fase de pré-processamento, com as probabilidades de cada uma destas pertencerem a cada uma das classes, e no fim do arquivo existe uma linha destinada a um "caso geral" que contém as probabilidades de uma palavra ao acaso ser de uma das C classes. Na Figura 4.1 são apresentados alguns exemplos de linhas do arquivo de mapeamento. Nesta pode-se perceber que há palavras que podem aparecer com mais do que uma classe (**finais**), sendo que algumas que possuem uma maior probabilidade de pertencer a algumas determinadas classes (**a**, por exemplo, tende a normalmente aparecer no *corpus* como o artigo definido feminino – etiqueta DET – e como uma preposição – etiqueta ADP. Todavia, este ainda pode aparecer como o nome da própria letra A – etiqueta N – ou como o pronome pessoal oblíquo feminino da 3ª

[1]Apesar dos numerais cardinais representarem um conjunto infinito de elementos, os nomes destes são compostos a partir de um conjunto finito de palavras. A última inovação notável na nomenclatura dos numerais cardinais data do século XVII, onde se criou a nomenclatura dos números maiores que 999.999.999, utilizando-se um radical latino seguido pela terminação **-lhão**. São exemplos desta nomenclatura: bilhão, quadrilhão, decilhão e centilhão.

	ADJ	ADP	ADV	CJ	DET	INTJ	MISC	MW	N	NUM	PART	PRON	PUNC	V
míope	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
esquadra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
finais	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
a	0.0	0.26	0.0	0.0	0.74	0.0	0.0	0.0	2e-4	0.0	0.0	4e-3	0.0	0.0
⋮														
-ano	0.45	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.55	0.0	0.0	0.0	0.0	0.0
-ém	0.0	0.0	0.77	0.0	0.0	0.0	0.0	0.0	0.06	0.0	0.0	0.14	0.0	0.03
-uém	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
⋮														
-	0.05	0.15	0.04	0.04	0.16	2e-4	4e-3	0.0	0.25	0.02	0.0	0.03	0.14	0.12

Figura 4.1: Exemplos de linhas do arquivo de mapeamento

pessoa do singular – etiqueta PRON. Neste últimos dois casos as probabilidades atribuídas a estas classes são ambas inferiores a 10^{-2}.). Também cabe notar que algumas terminações podem ser atribuídas a palavras de diversas classes, mas o acréscimo de uma letra nestas terminações podem reduzir drasticamente a gama de possíveis classes, facilitando o processo de desambiguação. Por exemplo, **-ém** pode ser um pronome em *alguém*, um verbo em *mantém*, um substantivo em *refém* e um advérbio em *além*, mas a terminação **-uém** somente é atribuída à classe de pronomes (*alguém*, *ninguém* etc)[2].

A escolha de quais terminações são relevantes para o etiquetador é determinada pelos quatro parâmetros abaixo discriminados:

- **Relevância das terminações** – este parâmetro é uma medida que determina o número mínimo de palavras distintas que devem terminar com uma sequência de caracteres específicos. Se uma dada terminação aparecer ao fim de ao menos este número de palavras, esta será considerada "relevante" para o processo de classificação gramatical. Esta medida foi pensada, pois há uma vasta gama de sequências de letras que não agregam qualquer informação relevante quanto à classe gramatical da palavra que esta compõe. Um exemplo de terminação que não agrega informação relevante quanto à classe gramatical de uma palavra é a *-vre* na palavra *livre*. Esta é um adjetivo, porém raríssimas são as palavras terminadas em *-vre*, determinando que esta não pode ser considerada uma terminação relevante para o processo de classificação;

- **Tamanho da menor terminação** – ele denota qual é o menor número de caracteres que uma terminação deve possuir para ser considerada uma candidata a ser usada no processo de classificação. A utilidade deste parâmetro

[2] A terminação **-uém** também está presente no advérbio *aquém*, mas devido ao seu raro uso, este não aparece no *corpus*, fazendo com que a terminação **-uém** seja, neste caso, encontrada somente em pronomes.

está em fazer com que o mWANN-Tagger evite empregar pequenas sequências de letras como terminações válidas para o procedimento classificatório. Isto é facilmente visto com a terminação *-e*, que está presente em muitas palavras (por conseguinte, seria aceitada pelo parâmetro supracitado), mas que não agrega informação quanto à classe de uma palavra;

- **Diferença entre o tamanho da maior e o da menor terminação** – este parâmetro foi criado para que se evite tanto a criação de dicionários enormes quanto o sobreajuste (*overfitting*) da rede. Optou-se por usá-lo, em vez do "tamanho da maior terminação", pois o primeiro pode assumir qualquer valor superior ou igual a 0, além de ser independente do "tamanho da menor terminação". O segundo, por outro lado, é dependente deste parâmetro e seu domínio de valores é variável, dependendo do valor que este tivesse;

- **Tamanho da menor "raiz"** – esta dissertação denota "raiz" como toda série de caracteres que não seja uma terminação. Este parâmetro é análogo ao "tamanho da menor terminação". Ele é usado pelo classificador, pois este precisa desconsiderar terminações cujo tamanho seja próximo ao da palavra à qual ele pertença. Um exemplo trivial é a palavra *sal* com a terminação *-al*. Esta costuma pertencer a adjetivos, tais como *legal*, *leal*, *constitucional*, dentre outros. *Sal*, no entanto, é um substantivo. O uso deste parâmetro, desta forma, evita que palavras como *sal* sejam classificadas incorretamente. Ademais, no inglês, o uso deste parâmetro não só pode evitar uma classificação incorreta, como pode auxiliar em se obter uma correta. A palavra *ugly*, por exemplo, que, apesar de possuir a terminação *-ly* (própria de advérbios) é um adjetivo. A ideia básica por trás disso é que o mWANN-Tagger, ao verificar que *ug-* é uma sequência de caracteres muito pequena, opte pela terminação *-y* (de adjetivo) em vez da *-ly*.

A linha de "caso geral" existente no arquivo de mapeamento é utilizada somente se existir alguma palavra que se deseja etiquetar e não existir nenhum vetor de probabilidades que possa ser atribuído a esta palavra. Este caso acontecerá se a palavra em questão não tiver aparecido durante a fase de treinamento e se nenhuma das terminações relevantes puder ser associada a esta palavra.

O passo seguinte da fase de pré-processamento consta na criação do arquivo de treinamento propriamente dito, que é montado com base no *corpus* original e nos dados coletados e inseridos no arquivo de mapeamento. O arquivo de treinamento é composto por uma série de linhas, cada uma associada a uma palavra do *corpus*. Em cada uma destas estão dispostas as probabilidades de cada palavra presente na **janela de contexto** da palavra associada a esta linha, juntamente com a classe

gramatical da palavra em questão, denotada por $C-1$ **0**s (as classes às quais a palavra desta linha não pertence) e um **1** (a classe à qual ela pertence). Cabe notar que deve ser dado um tratamento especial para palavras em início e fim de sentenças. Se estas não tiverem palavras suficientes em sua vizinhança para preencher sua janela de contexto, o espaço restante da janela deve ser preenchido com **0**s. Desta forma, cada linha do arquivo de treinamento é composta por $(A+D+2)\times C$ valores, sendo as C probabilidades de cada um das $A+D+1$ palavras da janela de contexto e mais C valores referentes à classe da palavra associada à linha em questão. A Figura 4.2 mostra como uma sentença transforma-se em uma série de linhas do arquivo de treinamento.

O último passo do pré-processamento das entradas consta na leitura de cada linha do arquivo de treino e a conversão dos primeiros $(A+D+1)\times C$ valores desta linha – as probabilidades das palavras na janela de contexto – em um vetor de *bits* para, enfim, utilizá-lo na entrada da rede WiSARD. Os últimos C valores das linhas são usados para determinar qual discriminador deve ser treinado, uma vez que estes valores determinam a classe da palavra sendo submetida ao treinamento.

A conversão das probabilidades em vetores de *bits* depende de um processo de discretização das mesmas. Utiliza-se um **índice de discretização** d, que representa o número de partes iguais nas quais o intervalo $[0,1]$ deve ser dividido. Obtendo-se, assim, d intervalos com o formato $\left(\frac{k}{d}, \frac{k+1}{d}\right], \forall k \in 0, ..., d-1$. O vetor de *bits* que representa uma probabilidade P discretizada em d *bits* é obtido se convertendo cada um dos d intervalos obtidos em um *bit* com o valor **1** se este abranger valor menores ou iguais a P e **0**, caso contrário. Um caso particular acontece se a probabilidade a ser convertida for $0,0$. Esta gerará um vetor de d *bits* possuindo **0** em todas as suas posições.

Na Figura 4.3 há um exemplo demonstrando o procedimento de discretização da probabilidade $P = 0,3$ em $d = 5$ *bits*. Neste caso, o intervalo $[0,1]$ é divido em intervalos de tamanho $0,2$. Neste caso, tanto o intervalo $[0;0,2]$ quanto o $[0,2;0,4]$ abrangem valores menores ou iguais a $0,3$. Desta forma, o resultado deste processo de discretização é o vetor de *bits* **11000**.

Possuindo as entradas já no formato de um vetor de *bits* e convertendo a saída para um vetor de *bits*, onde a probabilidade $1,0$ deva ser representada pelo *bit* **1** e as probabilidades $0,0$ pelo *bit* **0**, executa-se o treinamento da WiSARD presente na arquitetura do mWANN-Tagger. Devido aos possíveis ruídos que podem existir por causa de palavras que possuam classes gramaticais ambíguas ou possíveis anotações erradas no *corpus*, opta-se pela utilização da WiSARD com a extensão da DRASiW para que, assim, possa-se usar a técnica do *bleaching* durante a fase de etiquetagem.

Baseando-se nos passos após a criação do arquivo de mapeamento, nota-se que quatro outros parâmetros são importantes para o bom funcionamento do

Sentença de exemplo: Ele se divertirá!
Aparência no *corpus* padronizado: [ele|PRON] [se|PRON] [divertirá|V] [!|PUNC]

Arquivo de mapeamento:

	ADJ	ADP	ADV	CJ	DET	INTJ	MISC	MW	N	NUM	PART	PRON	PUNC	V
⋮														
ele	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
⋮														
se	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.86	0.0	0.0
⋮														
!	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
⋮														
-irá	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
⋮														

Linhas a serem transcodificadas:

	ADJ	ADP	ADV	CJ	DET	INTJ	MISC	MW	N	NUM	PART	PRON	PUNC	V
–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ele	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
se	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.86	0.0	0.0
PRON	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
ele	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
se	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.86	0.0	0.0
diverti**rá**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
PRON	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
se	0.0	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.86	0.0	0.0
diverti**rá**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
!	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
diverti**rá**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
!	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PUNC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Figura 4.2: Criação das linhas do arquivo de treinamento

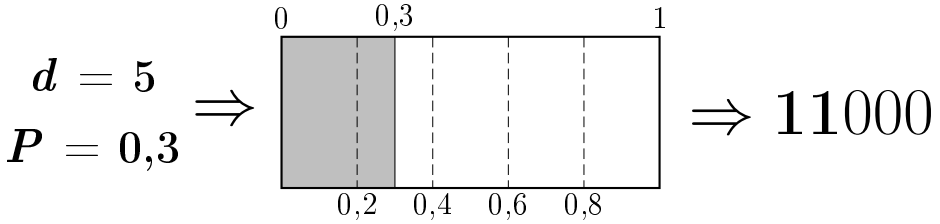


Figura 4.3: Processo de discretização de uma probabilidade

etiquetador. São eles:

- **Quantidade de palavras antes da atual** – este é um dos dois parâmetros que compõem a **janela de contexto**. Janelas muito pequenas costumam carecer de informações importantes para o processo de classificação. Por outro lado, as muito grandes podem ser danosas a este procedimento, uma vez que isto o torna mais suscetível a sobreajuste;
- **Quantidade de palavras depois da atual** – é o outro parâmetro que compõe a janela de contexto;
- **Índice de discretização** – como descrito acima, este índice é responsável por converter as probabilidades provenientes do arquivo de treino para um vetor de *bits*. Este parâmetro é necessário, pois o mWANN-Tagger usa a arquitetura WiSARD e, desta forma, somente pode receber entradas booleanas;
- **Número de *bits* de entrada das RAMs** – este parâmetro é responsável por determinar quão generalizante será o classificador. Quão menor for este número maior será a capacidade de generalização do classificador, e vice-versa.

Desta forma, necessita-se que se calibre corretamente oito parâmetros para que o mWANN-Tagger obtenha o máximo de exatidão em seu processo de etiquetagem. Como mencionado no início deste capítulo, espera-se que línguas isolantes sejam capazes de desambiguar as categorias gramaticais das palavras baseando-se no contexto onde esta está inserida e que as sintéticas seriam capazes de determinar a classe gramatical de uma palavra não apresentada no treinamento baseando-se na terminação que esta possuísse. Partindo-se desta hipótese, é esperado que línguas cujos índices de síntese sejam mais baixos requiram que os dois parâmetros que compõem a janela de contexto possuam um valor elevado. Por outro lado, as línguas sintéticas tenderiam a usar uma janela contexto menor e a optar pelo uso de uma maior gama de terminações para classificar palavras até então não apresentadas ao etiquetador.

O aprendizado do mWANN-Tagger, no entanto, não é somente neural, mas sim neurossimbólico. O etiquetador não executa o treinamento de palavras não ambíguas. Dada a característica determinística do funcionamento da rede WiSARD e da possibilidade de extração das informações aprendidas nesta, a partir de um processo de engenharia reversa, cogita-se que se podem extrair regras a partir do conhecimento inserido nas RAMs da rede. A única regra que se poderia obter com o treinamento de palavras não ambíguas seria que se esta sempre aparece com a mesma classe, somente pode ser classificada com esta mesma classe. Desta forma, sem treinar estes casos, tanto o treinamento do etiquetador quanto a etiquetagem

em si tornam-se mais rápidos e as únicas regras possíveis de serem extraídas da rede seriam as não triviais.

4.3 Etiquetagem

O procedimento de etiquetagem com o mWANN-Tagger funciona de forma análoga ao de treinamento. Neste, ao se apresentar ao mWANN-Tagger uma sentença para ser etiquetada gramaticalmente, ele transcodificará cada palavra da sentença em um vetor de probabilidades, baseando-se nas informações encontradas no arquivo de mapeamento. Feito isto, o etiquetador cria grupos de $(A+D+1) \times C$ probabilidades (as A palavras antes na janela de contexto, as D depois e a que se deseja etiquetador, cada uma destas representada por C probabilidades), executa o processo de discretização destas probabilidades e insere o vetor de *bits* proveniente deste processo na entrada da rede WiSARD. A etiqueta que será atribuída à palavra sendo submetida à classificação será aquela cujo discriminador retornar a maior medida de similaridade.

Como mencionado na seção anterior, o mWANN-Tagger utiliza a WiSARD com a extensão da DRASiW em seu treinamento. Faz-se uso desta extensão, pois necessita-se da técnica do *bleaching* durante a fase de etiquetagem, uma vez que há uma certa possibilidade de se obter dados ruidosos. Devido à grande gama de dados treinados, a escolha do limiar a ser usado no *bleaching* é feita utilizando a técnica *b-bleaching*.

Capítulo 5

Experimentação

Para se poder avaliar a função empregada por cada índice de síntese na tarefa de classificação gramatical, precisa-se descobrir quais parâmetros do mWANN-Tagger são diretamente afetados por uma simples mudança de valor em qualquer um destes índices. Para tal, serão usadas línguas que sejam razoavelmente diferentes entre si, para que, desta forma, possa-se obter conclusões satisfatórias. Para cada uma delas testou-se exaustivamente o mWANN-Tagger, utilizando uma vasta gama de configurações de parâmetros, até que se obtivesse a que produzisse os maiores valores de exatidão. As línguas escolhidas para os experimentos, assim como seus *corpora*, estão dispostas na seção abaixo.

5.1 Conjuntos de dados (*corpora*)

Oito línguas foram escolhidas para se usar nos experimentos, com o objetivo de se conseguir uma quantidade suficiente de dados para serem avaliados. Estas línguas estão listadas abaixo, incluindo as razões pelas quais se escolheu cada uma delas. Seus índices de síntese estão dispostos na Tabela 5.1, e o processo detalhado de como estes foram adquiridos encontra-se no apêndice A.

Tabela 5.1: Índices of síntese

Língua	Índices de síntese			
	Geral	Composicional	Derivacional	Inflexional
ZH	$1,537 \pm 0,058$	$1,537 \pm 0,058$	$0,000$	$0,0000$
EN	$1,394 \pm 0,075$	$1,043 \pm 0,024$	$0,244 \pm 0,059$	$0,108 \pm 0,036$
JA	$1,646 \pm 0,065$	$1,310 \pm 0,049$	$0,121 \pm 0,033$	$0,215 \pm 0,044$
PT	$1,898 \pm 0,111$	$1,082 \pm 0,031$	$0,214 \pm 0,052$	$0,602 \pm 0,102$
IT	$2,122 \pm 0,105$	$1,087 \pm 0,031$	$0,221 \pm 0,053$	$0,814 \pm 0,086$
DE	$1,978 \pm 0,111$	$1,078 \pm 0,032$	$0,377 \pm 0,078$	$0,522 \pm 0,071$
RU	$2,237 \pm 0,151$	$1,043 \pm 0,024$	$0,405 \pm 0,098$	$0,789 \pm 0,092$
TR	$2,072 \pm 0,151$	$1,123 \pm 0,042$	$0,220 \pm 0,056$	$0,729 \pm 0,137$

- **Mandarim** (código ISO-639-1: ZH – **Zh**ōngwén) – Foi escolhido pois é uma língua muito isolante. Morfemas derivacionais são muito raros e não existem inflexionais. Quase toda palavra nesta língua é composta puramente de raízes (normalmente uma ou duas);

- **Inglês** (código ISO-639-1: EN – **En**glish) – Pertence ao ramo germânico das línguas indo-europeias, o qual mantém tanto as construções composicionais quanto o sistema inflexional característicos desta macrofamília. No entanto, a língua inglesa simplificou drasticamente seu sistema inflexional com o passar dos anos (a sua conjugação verbal não depende mais nem de pessoa nem de número, o sistema de casos gramaticais deixou de existir, dentre outros). Ademais, esta língua importou uma grande gama de palavras do francês, tais como *table* e *independence*. A latinização de seu vocabulário contribuiu para que o índice de síntese derivacional do inglês crescesse. Devido ao seu sistema inflexional bastante simplificado, a possuir uma estrutura composicional simples e à influência indireta que esta obteve das regras derivacional provenientes do latim, a língua inglesa foi escolhida como uma das que serão usadas nos experimentos;

- **Japonês** (código ISO-639-1: JA) – Pertence à família japônica, a qual compreende somente um pequeno grupo de línguas, sendo o japonês a mais falada delas. Por tal razão, esta língua tem estruturas morfológicas muito particulares. No entanto, devido a influências de povos chineses, a língua japonesa antiga importou um série de palavras do chinês pré-médio, as quais hoje em dia compõem seu vocabulário sino-japonês. Estas palavras são compostas somente por raízes (normalmente uma ou duas), fazendo com que o índice composicional do japonês se tornasse relativamente alto, mesmo esta língua não sendo isolante como as chinesas;

- **Português** (código ISO-639-1: PT) – Dada a sua origem românica, esta língua raramente faz uso de composições. O latim, ao abolir as estruturas composicionais, as substituiu por novas formas de criação de palavras. Desta forma, foram introduzidos novos prefixos e sufixos derivacionais. Tal fato faz com que todas as línguas românicas tenham um índice de síntese composicional baixo e um derivacional razoavelmente alto. Apesar disso, a língua portuguesa não possui um índice composicional tão baixo, uma vez que esta faz um certo uso de contrações. Por tais motivos, e também por tanto ter simplificado a estrutura declinativa do latim quanto ter reduzido sua marcação de plural a uma simples estrutura aglutinativa (*-s*), que o português foi escolhido para fazer parte dos experimentos;

- **Italiano** (código ISO-639-1: IT) – Assim como a língua portuguesa, o italiano possui um índice de síntese composicional baixo e um derivacional razoavelmente alto. Porém, diferentemente desse, sua forma de plural manteve a estrutura fusional do latim. Tal estrutura também se manteve em outras classes gramaticais, dado que algumas elisões sonoras que ocorreram no português não afetaram o italiano (comparar *fiore*/*fiori* com *flor*/*flores*). A língua italiana também é caracterizada por possuir uma gama de contrações muito grande, bem superior à língua portuguesa, desta forma possuindo um índice composicional ligeiramente superior ao do português. Concluindo, faz-se uso do italiano, pois esta é uma língua românica que mantém o sistema fusional de inflexão do latim não só nos verbos quanto também nos substantivos e adjetivos;

- **Alemão** (código ISO-639-1: DE – **De**utsch) – Como um membro do ramo germânico das línguas indo-europeias, o alemão utiliza composição com uma de suas técnicas de formação de palavras. Ele também faz um uso intenso de derivações e seu sistema de inflexão não foi simplificado como o do inglês. A língua alemã aparece como uma boa candidata a ser usada nos experimentos, uma vez que esta representa as línguas do ramo germânico que mantiveram a maioria das técnicas de construção de palavras. Ademais, o alemão, diferentemente do inglês, não sofreu um processo massivo de latinização de seu vocabulário;

- **Russo** (código ISO-639-1: RU) – De todas as línguas indo-europeias listadas, esta é a única que manteve a maioria das técnicas de construção de palavras indo-europeias. Composições são razoavelmente comuns em russo, afixos derivacionais são usados em exaustão e seu sistema inflexional é grande e bem complexo. Esta língua está entre as escolhidas para os experimentos pois representa o conjunto de línguas fusionais bastante sintéticas;

- **Turco** (código ISO-639-1: TR) – Comumente conhecido como uma língua altamente aglutinativa. Turco, para contrabalancear a estrutura fusional russa, é usado nos experimentos representando o grupo de línguas que são muito sintéticas e puramente aglutinativas.

Um *corpus* anotado foi usado para cada língua (a informação detalhada deles está disposta na Tabela 5.2). O modelo WiSARD não requer que uma quantidade massiva de dados seja treinada nele, então, desta forma, somente 5000 sentenças foram escolhidas ao acaso de cada um destes *corpora* (exceto o *corpus* da língua italiana, o TUT (Turin University Treebank), que possui 2738 sentenças). Com estas foram criados *corpora* menores para cada uma das línguas.

Tabela 5.2: Os *corpora* usados nos experimentos

Língua	*Corpus*	Número de sentenças	Palavras por sentença
ZH	Penn Chinese Treebank 6.0 [56]	44517	24,70
EN	Brown Corpus [57]	57335	20,34
JA	TüBa-J/S – Tübinger Baumbank des Japanischen Spontansprache [58]	17626	8,90
PT	Bosque (Floresta Sintá(c)tica) [59]	9349	22,89
IT	TUT – Turin University Treebank [60]	2738	28,10
DE	NEGRA Corpus [61]	19235	17,57
RU	Conjunto de sentenças extraído do Russian National Corpus [62]	11237	14,80
TR	METU – Sabancı Turkish Treebank [63]	5623	9,60

Durante o procedimento de seleção de sentenças, fez-se a conversão dos conjuntos de classes originais de cada língua para o conjunto universal, mencionado na seção 4.1. Desta forma, as bases de dados usadas nos experimentos foram significativamente normalizadas, uma vez que estas tinham o mesmo conjunto de classes gramaticais e quase a mesma quantidade de sentenças.

5.2 Experimentos e análises

Uma série de experimentos foram feitos para testar diversos aspectos do etiquetador. Fez-se um para averiguar se há uma relação entre os índices de síntese das linguagens e a melhor configuração de parâmetros do mWANN-Tagger, e alguns outros para comparar seu desempenho, sua exatidão e sua precisão com sistemas que usam abordagens iterativas.

5.2.1 Função de cada índice de síntese em tarefas de classificação gramatical

Na Tabela 5.3 estão dispostos os melhores parâmetros para cada instância do mWANN-Tagger, assim como a porcentagem média de respostas exatas e o desvio padrão desta medida de exatidão. Como a quantidade de dados ainda é muito pequena, não há como se estimar uma função matemática propriamente dita destes. No entanto, algumas regras podem ser extraídas destes dados e, desta forma, descobrir-se o papel que cada índice de síntese emprega no procedimento de classificação gramatical utilizando redes neurais sem peso.

Para se descobrir se há uma dependência entre os índices de síntese e os valores da melhor configuração de parâmetros do etiquetador, extraiu-se o coeficiente de correlação de Pearson entre estas variáveis (ver Tabela 5.4 e Figura 5.1). A

Tabela 5.3: Melhores configurações do mWANN-Tagger para cada língua

Língua	Índice de síntese médio	Índice de síntese composicional médio	Índice de síntese derivacional médio	Índice de síntese inflexional médio	Índice de relevância das terminações	Tamanho da menor terminação	Diferença entre o tamanho da maior e o da menor terminação	Tamanho da menor "raiz"	Janela de contexto	Índice de discretização das probabilidades	Número de *bits* de entrada das RAMs	Porcentagem média de respostas exatas	Desvio padrão da exatidão das respostas
ZH	1,537	1,537	0,000	0,000	5	4	0	5	[1, 1]	86	37	94,36%	0,28%
EN	1,394	1,043	0,244	0,108	1	3	3	2	[1, 1]	44	38	97,86%	0,13%
JA	1,646	1,310	0,121	0,215	6	4	1	6	[1, 1]	52	49	98,94%	0,21%
PT	1,898	1,082	0,214	0,602	2	1	3	4	[1, 1]	27	36	98,30%	0,19%
IT	2,122	1,087	0,221	0,814	4	2	3	3	[1, 1]	35	35	98,19%	0,69%
DE	1,978	1,078	0,377	0,522	1	2	9	2	[1, 1]	46	34	97,70%	0,26%
RU	2,237	1,043	0,405	0,789	3	2	5	2	[1, 1]	36	31	98,32%	0,43%
TR	2,072	1,123	0,220	0,729	1	4	10	2	[0, 1]	11	19	97,13%	0,18%

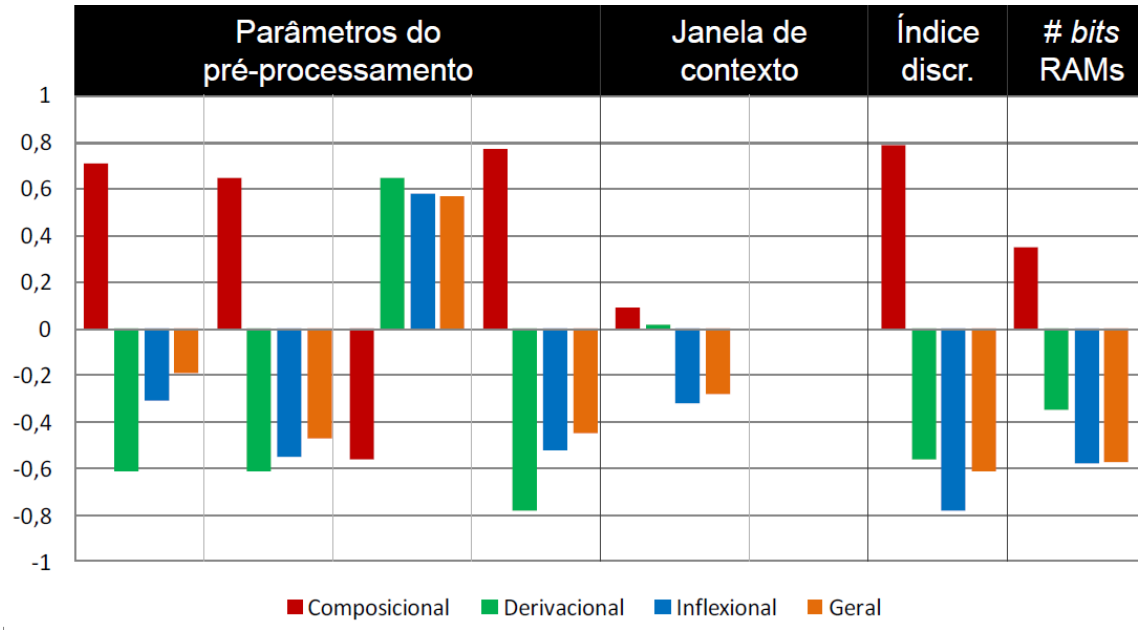


Figura 5.1: Comparação entre os coeficientes de correlação dos índices de síntese

Tabela 5.4: Coeficientes de correlação entre os índices de síntese das línguas e os valores dos parâmetros do mWANN-Tagger

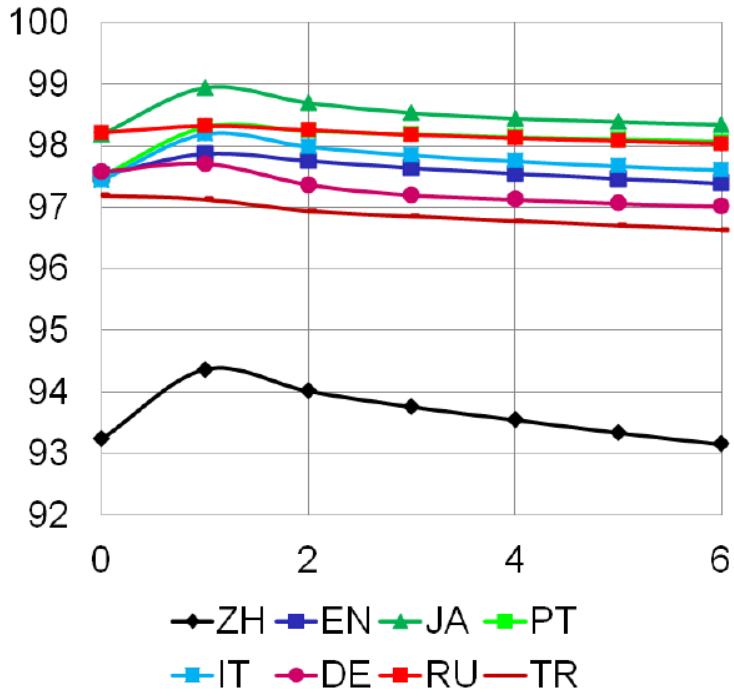
Tipo de índice	Índice de relevância das terminações	Tamanho da menor terminação	Diferença entre o tamanho da maior e o da menor terminação	Tamanho da menor "raiz"	Número de palavras antes (janela de contexto)	Número de palavras depois (janela de contexto)	Índice de discretização das probabilidades	Número de *bits* de entrada das RAMs
Composicional	0,71	0,65	-0,56	0,77	0,09	N/A	0,79	0,35
Derivacional	-0,61	-0,61	0,65	-0,78	0,02	N/A	-0,56	-0,35
Inflexional	-0,31	-0,55	0,58	-0,52	-0,32	N/A	-0,78	-0,58
Geral	-0,19	-0,47	0,57	-0,45	-0,28	N/A	-0,61	-0,57

partir destes dados, pode-se agrupar os parâmetros da rede em quatro categorias: *a*) aqueles que possuem uma correlação razoavelmente alta com, no mínimo, um dos índices (onde se enquadram o índice de relevância das terminações, o tamanho da menor "raiz" e o índice de discretização das probabilidades); *b*) os que possuem uma correlação moderada com a maioria dos índices (ambos parâmetros relacionados ao tamanho das terminações enquadram-se neste grupo); *c*) os que possuem uma correlação moderada a baixa com a maioria dos índices, mas ainda tem uma correlação moderada com um dos três índices (o número de *bits* de entradas nas RAMs pertence a este grupo), e *d*) aqueles cujos valores de correlação são baixos com todos os índices (onde se enquadra a janela de contexto).

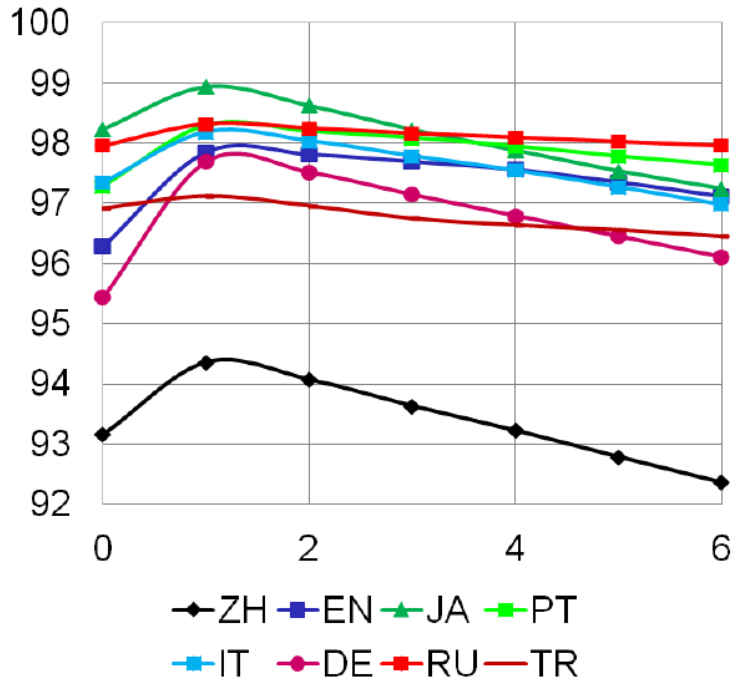
Pode-se perceber também que parâmetros que são diretamente proporcionais ao índice composicional tendem a ser inversamente proporcionais aos outros dois índices e vice-versa. Fato que também pode ser percebido quando compara-se os coeficientes de correlação de cada um dos três índices com os do índice de síntese geral. Neste último, alguns coeficientes de correlação foram mais baixos que em todos os outros índices, por exemplo, o coeficiente obtido quando feita a correlação deste índice com o parâmetro *índice de relevância das terminações* ou com o *tamanho da menor "raiz"*. Os outros coeficientes de correlação do índice de síntese geral normalmente são iguais ou próximos ao coeficiente mais baixo obtido por todos os índices.

Cabe notar também que o coeficiente de correlação dos parâmetros da janela de contexto são muito baixos, pois esta tende a ter um tamanho fixo, $[1, 1]$ (ver Tabela 5.3). Apesar desta ter o tamanho $[0, 1]$ para a língua turca, aumentar o número de palavras antes da atual para 1 quase não afetará a porcentagem de respostas exatas do classificador (97,13% para a configuração com a janela de contexto $[0, 1]$ contra 97,05% para a com a janela $[1, 1]$, como pode ser percebido na Figura 5.2a). Esta alteração irá, no entanto, prejudicar um pouco o desempenho deste, uma vez que o tamanho de entrada da rede aumentará. Um aumento deste no tamanho da janela de contexto faz com que o treinamento demore quase 50% a mais e a etiquetagem 25%.

Como pode ser visto na Figura 5.2, apesar da predileção pela janela de tamanho $[1, 1]$, tanto a língua turca quanto a russa (línguas bem sintéticas) não apresentam uma queda notável nas suas taxas de acerto caso o tamanho da janela de contexto seja alterado. Por outro lado, o mandarim, que é uma língua isolante, apresenta uma queda visível caso o tamanho da janela de contexto seja alterado. Esta queda na taxa de acerto é bem mais acentuada se algum dos parâmetros for zerado, fazendo com que o etiquetador somente considere o contexto proveniente de algum dos lados da palavra submetida à etiquetagem, e não de ambos. Comparando-se a Figura 5.2a com a Figura 5.2b nota-se também que a exatidão do etiquetador é mais sensível ao contexto presente após (ou à direita) da palavra a ser etiquetada do que o contexto



(a) Número de palavras antes



(b) Número de palavras depois

Figura 5.2: Impacto do tamanho da janela de contexto na taxa de acerto

Tabela 5.5: Tempo gasto pelo mWANN-Tagger em cada sentença, durante cada fase

Língua	Pré-processamento	Treinamento	Etiquetagem
ZH	104,420 ms	36,351 ms	114,337 ms
EN	51,386 ms	5,408 ms	14,522 ms
JA	4,142 ms	4,042 ms	9,186 ms
PT	54,823 ms	4,717 ms	13,003 ms
IT	17,519 ms	5,330 ms	15,016 ms
DE	89,843 ms	7,823 ms	22,055 ms
RU	75,819 ms	2,821 ms	7,197 ms
TR	38,641 ms	0,491 ms	1,600 ms

presente antes (ou à esquerda) desta.

5.2.2 Desempenho do mWANN-Tagger

Após feita a calibragem dos parâmetros do mWANN-Tagger, de forma a se ter a maior taxa de acerto, testou-se o etiquetador 200 vezes com esta configuração. Fez-se isso para se obter um valor bastante preciso de seu desempenho. A Tabela 5.5 e a Figura 5.3 contêm o tempo gasto pelo mWANN-Tagger em cada sentença, durante as fases de pré-processamento, treinamento e etiquetagem.

O tempo gasto pelo classificador em cada fase depende dos valores dos seus parâmetros. Levando-se em consideração que durante a fase de pré-processamento, o mWANN-Tagger cria tanto o arquivo de mapeamento quanto o de treinamento, qualquer parâmetro que afete o tamanho destes arquivos fará com que o tempo gasto nesta fase cresça (ou diminua).

Os parâmetros que são diretamente proporcionais ao tamanho de ambos os arquivos e, desta forma, do tempo gasto nesta fase, são a *diferença entre o tamanho da maior e o da menor terminação* (afetando o tamanho do arquivo de mapeamento) e ambos os parâmetros que mudam o tamanho da janela de contexto (alterando o tamanho do arquivo de treino). Os que são inversamente proporcionais são o *índice de relevância das terminações* e o *tamanho da menor "raiz"*, ambos sendo responsáveis pelo tamanho do arquivo de mapeamento. Contraintuitivamente, o *tamanho da menor terminação* não faz com que o tempo gasto nesta fase cresça nem decresça, uma vez que a *diferença entre o tamanho da maior e o da menor terminação* mantém a quantidade de terminações válidas praticamente idêntica.

Os tempos gastos tanto na fase de treinamento quanto na de etiquetagem dependem muito do tamanho da entrada da rede, o qual é diretamente proporcional ao *índice de discretização das probabilidades* e aos parâmetros da janela de contexto. O primeiro faz com que cada probabilidade seja convertida em um número maior de *bits* e o segundo aumenta a quantidade de palavras, e, por conseguinte, de

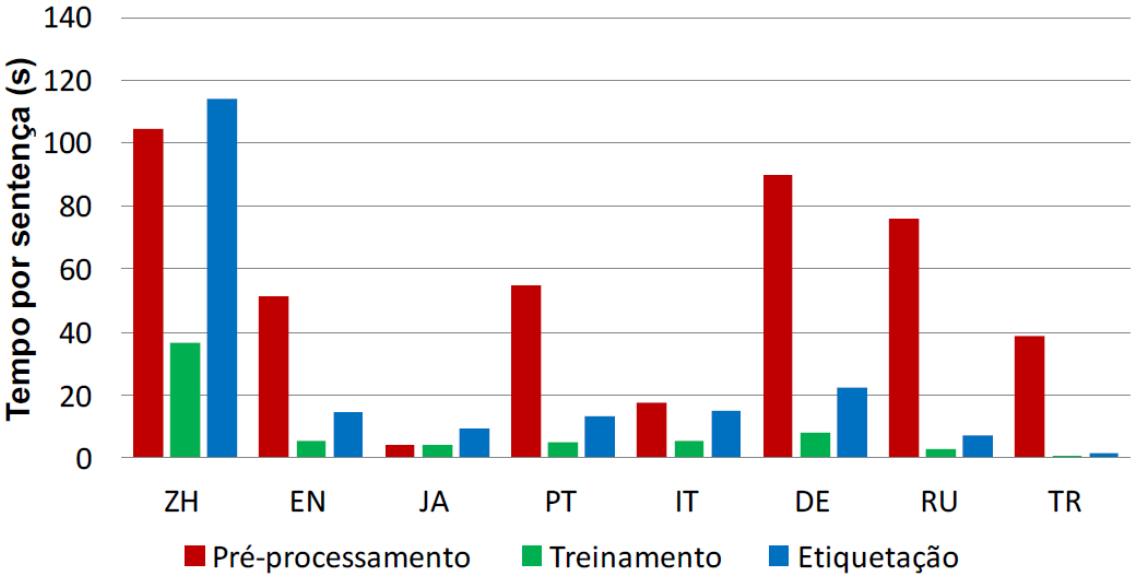


Figura 5.3: Comparação entre os tempos gastos pelo etiquetador em cada fase

probabilidades, na janela de contexto.

Considerando-se que o arquivo de mapeamento precisa ser acessado durante a fase de classificação, quaisquer parâmetros que afetem o seu tamanho da mesma forma o farão com o tempo gasto nessa fase. Todavia, seu efeito no desempenho do classificador nesta fase continua sendo muito pequeno, quando comparado com os três parâmetros mencionados anteriormente.

Tendo um tempo de treinamento baixo e a habilidade de aprender novos exemplos sem ter de reiniciar seu processo de aprendizado, o modelo WiSARD faz com que o mWANN-Tagger possa aprender novas sentenças em tempo de execução. Para fazê-lo, cada uma delas deve ser transcodificada em uma série de vetores contendo $(A + D + 2) \times C$ probabilidades, cujos valores devem ser extraídos do arquivo de mapeamento (ver a Seção 4.2 para compreensão da quantidade de probabilidades nos vetores). Após transcodificadas, estas probabilidades devem ser convertidas para vetores de *bits*, para, então, estes serem treinados. A integridade dos dados já aprendidos é garantida, uma vez que o treinamento da WiSARD requer uma simples operação de escrita em memória.

5.2.3 Comparação da taxa de acerto com o estado da arte

PETROV *et al.* [1] criou um conjunto universal de classes gramaticais e implementou um classificador gramatical *Trigrams'n'Tags* (*TnT POS-Tagger*) – um classificador que faz uso de modelos ocultos de Markov baseado em trigramas – para poder testá-lo. O propósito do trabalho deles era mostrar como um conjunto universal de classes poderia auxiliar na produção de melhores resultados em tarefas de classificação gramatical. O mWANN-Tagger também é um classificador que faz

Tabela 5.6: Comparação entre as taxas de acerto obtidas pelo mWANN-Tagger e pelo *TnT-Tagger* empregado em [1]

Língua	mWANN-Tagger	PETROV *et al.* [1]	Usa o mesmo *corpus*?
ZH	**94,36%**	93,4%	Sim
EN	**97,86%**	96,8%	Não
JA	**98,94%**	98,0%	Sim
PT	**98,30%**	96,8%	Sim
IT	**98,19%**	95,8%	Não
DE	97,70%	**97,9%**	Sim
RU	**98,32%**	96,8%	Sim
TR	**97,13%**	89,1%	Sim

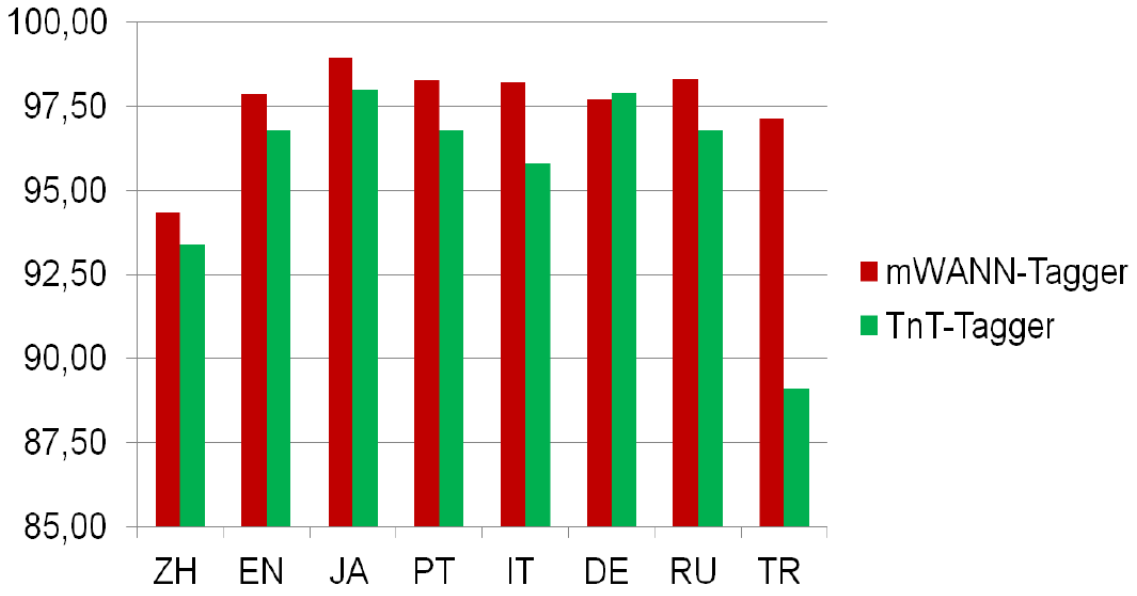


Figura 5.4: Comparação entre as taxas de acerto do mWANN-Tagger com as obtidas por um modelo estocástico

uso de um conjunto de classes gramaticais padronizado, de forma a poder utilizá-lo para qualquer língua.

Nesta seção as taxas de acerto do mWANN-Tagger serão comparadas com as obtidas em [1]. Esta comparação pode ser feita através da análise dos resultados apresentados na Tabela 5.6 e na Figura 5.4. Os melhores resultados de cada língua estão marcados em **negrito**. Os resultados de PETROV *et al.* [1] que serão usados nesta comparação serão aqueles que, tal como o mWANN-Tagger, usam o conjunto universal de classes gramaticais tanto durante a fase de treinamento quanto durante a de etiquetagem.

De acordo com a Tabela 5.6[1], as taxas de acerto do mWANN-Tagger são maiores que as obtidas em [1]. A única exceção, a taxa de acerto para a língua alemã, foi somente 0,2% menor. Ademais, os desvios padrões dos valores de exatidão do mWANN-Tagger eram menores que 0,7%, enquanto que os obtidos por PETROV *et al.* [1], utilizando o *TnT-Tagger*, eram geralmente em torno de 2,25%.

5.2.4 mWANN-Tagger *versus* classificador neural do tipo *feedforward*

Por fim, optou-se por comparar o mWANN-Tagger com uma abordagem cognitiva usando um modelo neural com pesos, uma rede *perceptron* multicamadas (*multilayer perceptron* – MLP). Neste caso, ambos modelos foram testados utilizando validação cruzada de 10 iterações (*10-fold cross-validation*). Empregou-se as mesmas bases dos experimentos anteriores (as que constam na seção 5.1). No entanto, testou-se também estes classificadores com versões menores destas bases (com 50, 108, 232, 500, 1077 e 2321 sentenças – $50 \times 10^{\frac{k}{3}}, k \in \{0,1,2,3,4,5\}$[2]).

Para a implementação do modelo *perceptron* multicamadas utilizou-se uma biblioteca para linguagem Java onde possuísse uma forma simples de instanciar uma rede deste tipo, e cujos resultados se mostrassem bons, a fim de se tentar superar os resultados obtidos pelo mWANN-Tagger. A biblioteca escolhida foi a Encog [64].

Experimentou-se, então, uma variedade de configurações para a rede *feedforward* a fim de se obter a que apresentasse os melhores resultados na maioria dos casos. A rede escolhida possui 7 neurônios na camada escondida, sua taxa de aprendizado é de $\eta = 0,001$, momento $\alpha = 0,1$, seu treinamento é baseado em épocas e possui três critérios de parada – *a*) chegar a 1000 épocas; *b*) o erro quadrático médio ser menor que 10^{-3}, e *c*) durante 5 épocas seguidas o erro quadrático médio ter uma alteração em seu valor menor que 10^{-6}. Se qualquer um destes critérios for satisfeito, o aprendizado da rede para. O tamanho de época usado foi aquele que compreendesse todos os dados de treinamento.

As redes foram comparadas em três aspectos, seus tempos de treinamento, de etiquetagem e suas taxas de acerto. Os resultados destas comparações estão dispostos nas Tabelas 5.7, 5.8 e 5.9, respectivamente[3]. Os tempos gastos na fase

[1] A base usada para língua russa foi a mesma usada em [1], no entanto, quando esta dissertação foi composta, somente um conjunto de sentenças aleatórias do *Russian National Corpus* estava disponível *off-line* gratuitamente. Devido a alguns problemas técnicos e/ou de *copyright*, o restante do *corpus* somente estava acessível *on-line*.

[2] Escolheu-se esta quantidade de amostras de treino para se poder averiguar o desempenho dos classificadores em amostras de tamanhos bem variados.

[3] O tamanho da amostra na Tabela 5.7 é igual à quantidade de sentenças usadas para treino em uma iteração, isto é 90% do total de sentenças da base usada (no caso de *10-fold cross-validation*). Analogamente, os tamanhos das amostras na Tabela 5.8 é igual à quantidade de sentenças usadas para classificação em uma iteração, isto é 10% do total.

Tabela 5.7: Tempo médio gasto na fase de treinamento × Número de sentenças

Tamanho da amostra	mWANN-Tagger		MLP	
	Menor tempo	Maior tempo	Menor tempo	Maior tempo
45	0,021s (TR)	0,303s (DE)	2,923s (RU)	4,803s (DE)
97	0,042s (TR)	1,347s (ZH)	6,604s (JA)	19,881s (IT)
209	0,052s (TR)	4,369s (ZH)	6,154s (TR)	39,827s (ZH)
450	0,087s (TR)	10,814s (ZH)	15,858s (TR)	58,581s (EN)
969	0,337s (TR)	47,181s (ZH)	42,478s (TR)	271,821s (EN)
2089	0,406s (TR)	61,755s (ZH)	17,855s (IT)	151,438s (JA)
4500	2,138s (TR)	165,649s (ZH)	4,791s (EN)	108,865s (DE)

Tabela 5.8: Tempo médio gasto na fase de etiquetagem × Número de sentenças

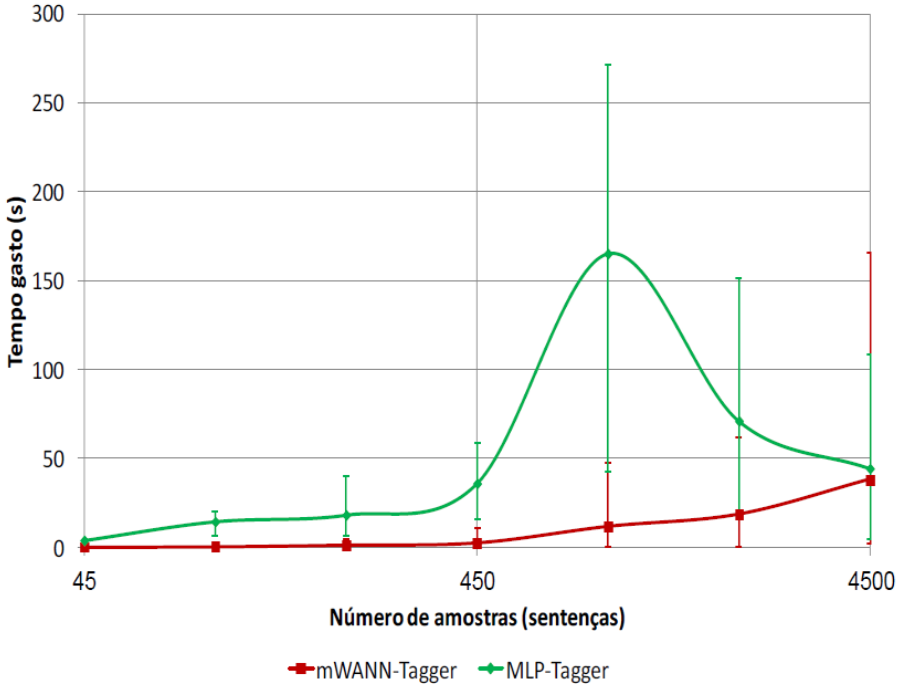
Tamanho da amostra	mWANN-Tagger		MLP	
	Menor tempo	Maior tempo	Menor tempo	Maior tempo
5	0,007s (RU)	0,083s (DE)	0,001s (PT)	0,001s (ZH)
11	0,013s (TR)	0,304s (ZH)	0,001s (JA)	0,005s (ZH)
23	0,012s (TR)	1,111s (ZH)	0,002s (JA)	0,007s (ZH)
50	0,027s (TR)	3,122s (ZH)	0,003s (JA)	0,016s (EN)
108	0,097s (TR)	14,367s (ZH)	0,010s (JA)	0,057s (IT)
232	0,122s (TR)	20,042s (ZH)	0,010s (TR)	0,057s (IT)
500	0,765s (TR)	57,269s (ZH)	0,037s (JA)	0,094s (IT)

de pré-processamento, por outro lado, não são comparados, uma vez que ambas as redes executam esta fase de forma idêntica, criando um arquivo de mapeamento e um de treino.

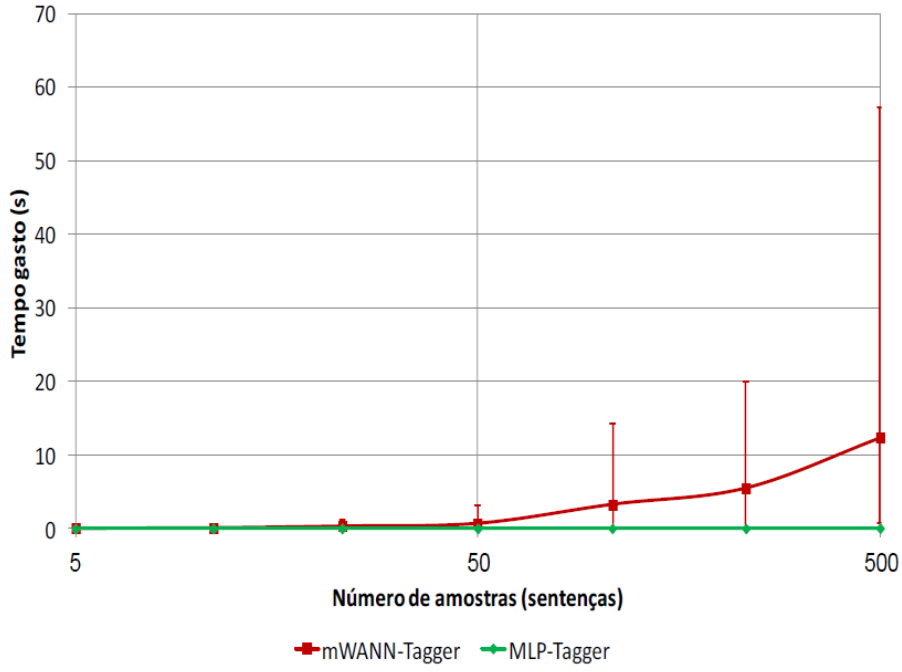
Inicialmente, pode-se notar pelos resultados das Tabelas 5.7 e 5.9 que a rede MLP apresentou um comportamento anômalo quando as amostras de treino possuem 969 ou mais sentenças (e, analogamente, quando o número de sentenças nas amostras de teste é superior ou igual a 108 e o número total de sentenças usadas é 1077). Nestes casos, a rede sofreu paradas prematuras durante sua fase de treinamento. Estas ocorreram pois houve cinco épocas seguidas onde as alterações do erro quadrático médio foram inferiores a 10^{-6}. Esta parada prematura pode ser mais facilmente

Tabela 5.9: Taxa de acerto média × Número de sentenças

Tamanho da amostra	mWANN-Tagger		MLP	
	Menor taxa de acerto	Maior taxa de acerto	Menor taxa de acerto	Maior taxa de acerto
50	92,07% (DE)	98,01% (JA)	71,06% (TR)	85,50% (EN)
108	95,23% (DE)	98,04% (RU)	74,44% (TR)	90,81% (EN)
232	94,69% (ZH)	97,77% (RU)	86,76% (ZH)	95,06% (EN)
500	93,89% (ZH)	98,05% (JA)	89,08% (ZH)	97,98% (RU)
1077	94,22% (ZH)	98,32% (JA)	73,85% (IT)	97,98% (RU)
2321	94,35% (ZH)	98,62% (JA)	10,35% (IT)	93,04% (JA)
5000	94,36% (ZH)	98,94% (JA)	1,90% (DE)	73,15% (TR)



(a) Tempo médio gasto na fase de treinamento



(b) Tempo médio gasto na fase de etiquetagem

Figura 5.5: Tempos gastos pelo mWANN-Tagger e por um etiquetador MLP

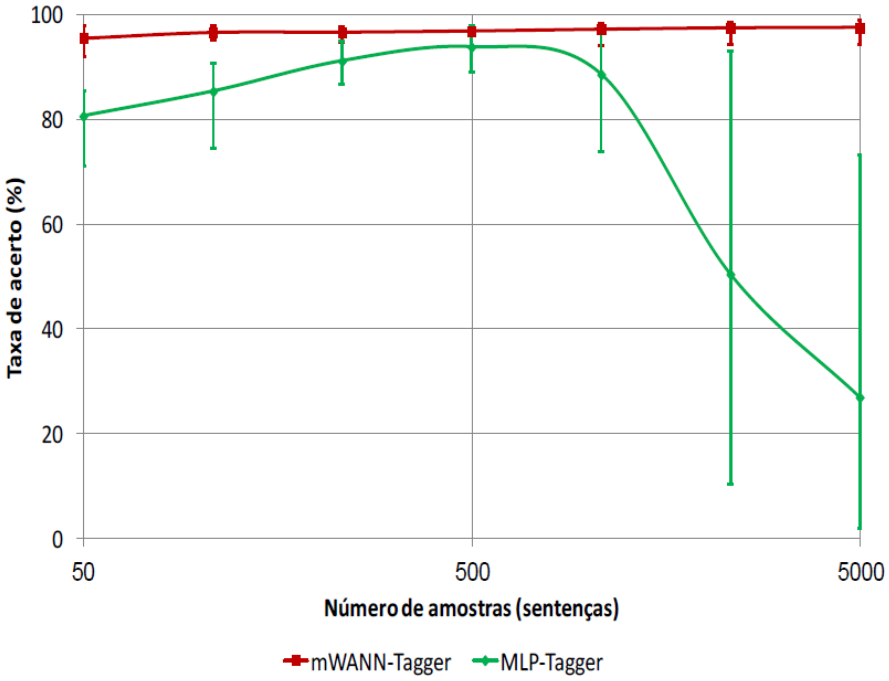


Figura 5.6: Comparação da taxa média de acerto do mWANN-Tagger com a de um etiquetador MLP

percebida nos gráficos da Figura 5.5 e na Figura 5.6. Nos casos restantes, a rede MLP somente parou o seu treinamento quando se atingiu 1000 épocas. As comparações apresentadas nos parágrafos seguintes não estarão levando em consideração os casos onde houve parada prematura no treinamento da rede MLP.

Os dados da Tabela 5.7 mostram que o mWANN-Tagger é superior à rede *perceptron* multicamadas durante a fase de treinamento, uma vez que este ou executa uma simples operação de escrita em memória (quando não tem plena certeza da classe da palavra) ou não executa operação alguma, caso a palavra em questão esteja constando no arquivo de mapeamento como associada a uma única classe. Este procedimento de escolha da classe baseado em um dicionário, apesar de ser inicialmente arriscado, torna-se bastante confiável conforme a quantidade de amostras de treinamento aumenta. O que, pela Tabela 5.9, comprova a capacidade do mWANN-Tagger, dado que ele pode ser treinado em bases de diversos tamanhos, pois este não apresenta efeitos de sobreajuste (*overfitting*) – fato minimizado pelo conjunto DRASiW + *bleaching*. O gráfico na Figura 5.5a mostra ainda que o tempo gasto durante o treino da rede MLP tende a aumentar vertiginosamente se o número de sentenças a serem treinadas crescer. Tal aumento somente não foi maior devido à parada prematura da rede MLP.

Ademais, mesmo em bases com pouca informação o mWANN-Tagger consegue ter uma boa exatidão em suas classificações. A Tabela 5.9 e a Figura 5.6 mostram que o seu resultado mais baixo (**92,07%** com a língua alemã) foi 6,5% melhor que

o mais alto obtido pela rede *perceptron* multicamadas (**85,50%** na língua inglesa). Outros experimentos também demonstram que o mWANN-Tagger tende a ter uma maior certeza em suas respostas, atingindo desvios padrões de 0,5% a 2,5% menores que a rede MLP (cujos desvios padrão são ainda assim menores que os obtidos em [1]).

A Figura 5.6 também mostra que a taxa média de acerto do mWANN-Tagger aumenta conforme o número de sentenças aprendido cresce. Isto significa que o mWANN-Tagger possivelmente pode vir a produzir taxas de acerto ainda maiores do que as já obtidas, caso se execute um treinamento com todas as sentenças dos *corpora* usados para estes experimentos. Desta forma, superando ainda mais os resultados obtidos em [1].

Vê-se também que o tempo de treinamento do mWANN-Tagger somente aumenta porque a quantidade de amostras também o faz. Há uma relação aparentemente linear entre o número de sentenças e o tempo gasto na fase de treinamento. O mesmo ocorre com a rede *perceptron* multicamadas, sendo que neste caso deve-se ao limite de 1000 épocas para se treinar a rede. Por outro lado, o tempo de classificação do mWANN-Tagger é bem maior que o obtido pela rede MLP. Mas tal fato pode ser compreendido como o tempo gasto para executar os desempates das respostas dos vários discriminadores, o qual é compensado pela alta taxa de acerto do classificador. Ademais, normalmente não se deseja classificar uma gama grande de sentenças ao mesmo tempo.

Capítulo 6

Conclusões

Nesta seção serão apresentadas as considerações finais e serão propostos alguns trabalhos futuros, tais como evoluções do mWANN-Tagger ou outras arquiteturas que possam usá-lo como base.

6.1 Considerações Finais

Este trabalho propôs o uso dos índices de síntese das linguagens como uma variável fundamental e suficiente para que se obtivesse as classes gramaticais das palavras corretamente. Para tal, se apresentou um etiquetador gramatical que utiliza uma arquitetura baseada na rede neural sem peso WiSARD, tal como também foi proposta uma técnica para se estimar a melhor configuração de parâmetros deste de forma mais prática e direta. Os benefícios da arquitetura sem peso e da técnica de estimação de parâmetros fizeram com que o mWANN-Tagger se mostrasse mais ágil que os classificadores existentes na literatura. Sua exatidão foi mensurada e se averigou que ele normalmente é superior ao estado da arte (ou equiparável, na língua alemã).

Na estimativa de parâmetros foram encontrados alguns resultados contraintuitivos. O mais notável destes foi o tamanho da janela de contexto, o qual é preferível que seja fixo em uma palavra antes do termo corrente e uma depois, ao invés de ser variável, tendendo a ser maior em línguas mais isolantes e menor em línguas mais sintéticas (ver hipóteses na Seção 1.1). Contudo, isto ainda pode ser verificado na língua turca cuja melhor configuração de parâmetros do mWANN-Tagger contou com uma janela de contexto sem nenhuma palavra antes da corrente.

6.2 Trabalhos futuros

Este classificador emprega as terminações e/ou, analogamente, os primeiros caracteres das palavras durante tanto seu procedimento de treinamento quanto de classificação. Isto impede que ele seja usado para classificar gramaticalmente palavras em alguma língua de morfologia não concatenativa, tais como árabe, hebraico, sírio, entre outras [65]. Estas línguas são caracterizadas por possuir um sistema de derivação que consta na alteração das vogais presentes no meio das palavras. Em árabe, por exemplo, o plural da palavra *kitāb* (livro) é *kutub* (livros). Estas possuem um radical triconsonantal *k-t-b*, que destes geram-se outras palavras como *maktūb* (está escrito). Desta forma, mostra-se muito interessante que se adicione ao mWANN-Tagger técnicas de linguística computacional orientadas a línguas semitas [66–68], de forma que este classificador seja capaz de classificar com precisão as palavras deste conjunto de línguas.

Outro possível aprimoramento para o mWANN-Tagger seria, baseado no trabalho de NASEEM *et al.* [38], transformá-lo em classificador gramatical capaz de executar tarefas de classificação de forma não supervisionada. Para tal, o mWANN-Tagger teria de ser capaz de usar informações interlinguais (*cross-lingual*), como feito no artigo previamente mencionado. Esta abordagem não supervisionada poderia ser implementada com o uso da AUTOWiSARD [46]. Este aprimoramento tornaria o mWANN-Tagger capaz de etiquetar sentenças em línguas polissintéticas, pois estas são pouco faladas e, por isso, não se encontram *corpora* anotados nestas línguas.

Por fim, o mWANN-Tagger poderia ser utilizado como base de um sistema de indução gramatical baseado em redes neural sem peso. Devido a seu alto desempenho tanto na fase de treinamento quanto na de classificação, este poderia ser usado facilmente em uma arquitura de aprendizado mútuo. Fato este que é corroborado pela capacidade do mWANN-Tagger poder aprender em tempo de execução e por este aparentemente não sofrer os efeitos de sobreajuste (*overfitting*). E considerando os benefícios que a indução gramatical pode oferecer às tarefas de tradução automática, como mostrado em [69], pode-se afirmar que a criação de um sistema assim criaria um elo importante entre as arquiteturas cognitivas para aquisição de linguagem e as tarefas de tradução automática. Ademais, como mencionado em [70], sistemas de indução de gramáticas podem ser usados nos mais diversos campos, não somente nos linguísticos. Entre eles, destacam-se a bioinformática e o tratameto de séries temporais.

Referências Bibliográficas

[1] PETROV, S., DAS, D., MCDONALD, R. "A Universal Part-of-Speech Tagset". In: *Proceedings of 8th International Conference on Language Resources and Evaluation (LREC)*, maio 2012.

[2] VAUQUOIS, B. "A survey of formal grammars and algorithms for recognition and transformation in machine translation". In: *IFIP Congress-68*, p. 254–260, Edinburgh, 1968.

[3] GREENBERG, J. H. "A quantitative approach to the morphological typology of language", *International Journal of American Linguistics*, v. 26, n. 3, pp. 178–194, 1960.

[4] HAKKANI-TÜR, D. Z., OFLAZER, K., TÜR, G. "Statistical morphological disambiguation for agglutinative languages". In: *Proceedings of the 18th Conference on Computational linguistics - Volume 1*, COLING '00, p. 285–291, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[5] PIRKOLA, A. "Morphological typology of languages for IR", *Journal of Documentation*, v. 57, n. 3, pp. 330–348, 2001.

[6] JURAFSKY, D., MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 2 ed. Upper Saddle River, NJ, USA, Prentice Hall, 2008.

[7] HARRIS, Z. S. *String Analysis of Sentence Structure*. The Hague, Mouton, 1962.

[8] KARLSSON, F., VOUTILAINEN, A., HEIKKILÄ, J., et al. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin, Mouton de Gruyter, 1995.

[9] KUPIEC, J. "Robust part-of-speech tagging using a hidden Markov model", *Computer Speech and Language*, v. 6, pp. 225–242, 1992.

[10] WEISCHEDEL, R., SCHWARTZ, R., PALMUCCI, J., et al. "Coping with ambiguity and unknown words through probabilistic models", *Computational Linguistics*, v. 19, n. 2, pp. 361–382, jun. 1993.

[11] MERIALDO, B. "Tagging English text with a probabilistic model", *Computational Linguistics*, v. 20, n. 2, pp. 155–171, jun. 1994.

[12] SCHÜTZE, H., SINGER, Y. "Part-of-speech tagging using a Variable Memory Markov model". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, p. 181–187, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[13] BRANTS, T. "TnT: A statistical part-of-speech tagger". In: *Proceedings of the 6TH conference on Applied Natural Language Processing*, ANLC '00, p. 224–231, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[14] MCCALLUM, A., FREITAG, D., PEREIRA, F. C. N. "Maximum Entropy Markov Models for Information Extraction and Segmentation". In: *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, p. 591–598, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[15] DENIS, P., SAGOT, B. "Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort". In: Kwong, O. (Ed.), *PACLIC*, p. 110–119. City University of Hong Kong Press, 2009.

[16] LAFFERTY, J. D., MCCALLUM, A., PEREIRA, F. C. N. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[17] SUTTON, C., MCCALLUM, A. "An introduction to conditional random fields for relational learning". In: Getoor, L., Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, MIT Press, Cambridge, MA, USA, 2007.

[18] CONSTANT, M., SIGOGNE, A. "MWU-aware part-of-speech tagging with a CRF model and lexical resources". In: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, p. 49–56, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[19] SCHMID, H. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[20] JELINEK, F., LAFFERTY, J. D., MAGERMAN, D. M., et al. "Decision tree parsing using a hidden derivation model". In: *Proceedings of the workshop on Human Language Technology*, HLT '94, p. 272–277, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[21] SCHMID, H. "Improvements In Part-of-Speech Tagging With an Application To German". In: *Proceedings of the ACL SIGDAT-Workshop*, p. 47–50, 1995.

[22] MAGERMAN, D. M. "Statistical decision-tree models for parsing". In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, p. 276–283, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[23] SCHMID, H. "Part-of-speech tagging with neural networks". In: *Proceedings of the 15th Conference on Computational linguistics - Volume 1*, COLING '94, p. 172–176, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[24] BRILL, E. "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging", *Computational Linguistics*, v. 21, n. 4, pp. 543–565, dez. 1995.

[25] DAELEMANS, W., ZAVREL, J., BERCK, P., et al. "MBT: A Memory-Based Part of Speech Tagger-Generator". In: *Proceedings of the 4th Workshop on Very Large Corpora*, p. 14–27. ACL SIGDAT, 1996.

[26] RATNAPARKHI, A. "A Maximum Entropy Model for Part-Of-Speech Tagging". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 133–142. Association for Computational Linguistics, abr. 1996.

[27] TSURUOKA, Y., TATEISHI, Y., KIM, J.-D., et al. "Developing a Robust Part-of-Speech Tagger for Biomedical Text". In: Bozanis, P., Houstis, E. N. (Eds.), *Proceedings of the 10th Panhellenic Conference on Informatics PCI-2005*, v. 3746, *Lecture Notes in Computer Science*, p. 382–392, Berlin, Heidelberg, nov. 2005. Springer-Verlag.

[28] TSURUOKA, Y., TSUJII, J. "Bidirectional inference with the easiest-first strategy for tagging sequence data". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 467–474, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[29] TOUTANOVA, K., KLEIN, D., MANNING, C. D., et al. "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[30] MANNING, C. D. "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" In: *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, p. 171–189, Berlin, Heidelberg, 2011. Springer-Verlag.

[31] GIMÉNEZ, J., MÀRQUEZ, L. "SVMTool: A general POS tagger generator based on Support Vector Machines". In: *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, p. 43–46, 2004.

[32] SØGAARD, A. "Semi-supervised condensed nearest neighbor for part-of-speech tagging". In: *ACL (Short Papers)*, p. 48–52, Stroudsburg, PA, USA, 2011. The Association for Computer Linguistics.

[33] COLLINS, M. "Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms". In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, p. 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[34] SPOUSTOVÁ, D., HAJIČ, J., RAAB, J., et al. "Semi-supervised training for the averaged perceptron POS tagger". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, p. 763–771, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[35] SHEN, L., SATTA, G., JOSHI, A. "Guided Learning for Bidirectional Sequence Classification". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 760–767, Stroudsburg, PA, USA, jun. 2007. Association for Computational Linguistics.

[36] SNYDER, B., NASEEM, T., EISENSTEIN, J., et al. "Unsupervised multilingual learning for POS tagging". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 1041–1050, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[37] SNYDER, B., NASEEM, T., EISENSTEIN, J., et al. "Adding more languages improves unsupervised multilingual part-of-speech tagging: a Bayesian non-parametric approach". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, p. 83–91, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[38] NASEEM, T., SNYDER, B., EISENSTEIN, J., et al. "Multilingual Part-Of-Speech Tagging: Two Unsupervised Approaches", *Journal of Artificial Intelligence Research*, v. 36, pp. 341–385, 2009.

[39] MA, Q., MURATA, M., UCHIMOTO, K., et al. "Hybrid neuro and rule-based part of speech taggers". In: *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, p. 509–515, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[40] MARQUES, N. C., BADER, S., ROCIO, V., et al. "Neuro-Symbolic Word Tagging". In: Neves, J., Santos, M. F., Machado, J. (Eds.), *New Trends in Artificial Intelligence*. APPIA - Associação Portuguesa para a Inteligência Artificial, dez. 2007.

[41] CARNEIRO, H. C. C., FRANÇA, F. M. G., LIMA, P. M. V. "WANN-Tagger: A Weightless Artificial Neural Network Tagger for the Portuguese Language". In: Filipe, J., Kacprzyk, J. (Eds.), *ICFC-ICNC 2010 - Proceedings of the International Conference on Fuzzy Computation and International Conference on Neural Computation*, p. 330–335. SciTePress, out. 2010.

[42] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2 ed. Upper Saddle River, NJ, USA, Prentice Hall, 1998.

[43] HORNIK, K., STINCHCOMBE, M., WHITE, H. "Multilayer feedforward networks are universal approximators", *Neural Networks*, v. 2, n. 5, pp. 359–366, jul. 1989.

[44] ALEKSANDER, I., DE GREGORIO, M., FRANÇA, F. M. G., et al. "A brief introduction to Weightless Neural Systems." In: *ESANN*, p. 299–305, 2009.

[45] ALEKSANDER, I., THOMAS, W. V., BOWDEN, P. A. "WISARD: A Radical Step Foward in Image Recognition", *Sensor Review*, v. 4, pp. 120–124, jul. 1984.

[46] WICKERT, I., FRANÇA, F. M. G. "AUTOWISARD: Unsupervised Modes for the WISARD". In: Mira, J., Prieto, A. (Eds.), *Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence*, v. 2084, *Lecture Notes in Computer Science*, Springer-Verlag, p. 435–441, Berlin, Heidelberg, 2001.

[47] KANERVA, P. *Sparse Distributed Memory*. Cambridge, MA, USA, MIT Press, 1988.

[48] KAN, W.-K., ALEKSANDER, I. "A Probabilistic Logic Neuron Network for Associative Learning". In: *IEEE First International Conference on Neural Networks*, v. 2, p. 541–548, out. 1987.

[49] DE BARROS CARVALHO FILHO, E. C., FAIRHURST, M. C., BISSET, D. L. "Adaptive pattern recognition using goal seeking neurons", *Pattern Recognition Letters*, v. 12, n. 3, pp. 131–138, 1991.

[50] ALEKSANDER, I. "Ideal Neurons for Neural Computers". In: Eckmiller, R., Hartmann, G., Hauske, G. (Eds.), *Parallel Processing in Neural Systems and Computers*, Elsevier Science Inc., New York, NY, USA, 1990.

[51] MRSIC-FLOGEL, J. "Convergence Properties of Self-Organizing Maps". In: Kohonen, T., Mäkisara, K., Simula, O., et al. (Eds.), *Proceedings of the International Conference on Artificial Neural Networks 1991* (Espoo, Finland), v. 1, p. 879–886. Amsterdam; New York: North-Holland, 1991.

[52] ALEKSANDER, I., MORTON, H. B. "General neural unit: retrieval performance", *Electronics Letters*, v. 27, n. 19, pp. 1776–1778, 1991.

[53] DE GREGORIO, M. *On the reversibility of multi-discriminator systems*. Relatório Técnico 125/97, Istituto di Cibernetica–CNR, 1997.

[54] SOARES, C. M., DA SILVA, C. L. F., DE GREGORIO, M., et al. "Uma implementação em software do classificador WISARD". In: *V Simpósio Brasileiro de Redes Neurais*, v. 2, p. 225–229, Belo Horizonte, MG, Brazil, dez. 1998.

[55] GRIECO, B. P. A., LIMA, P. M. V., DE GREGORIO, M., et al. "Producing pattern examples from "mental" images", *Neurocomputing*, v. 73, n. 7–9, pp. 1057–1064, mar. 2010.

[56] XIA, F., PALMER, M., XUE, N., et al. "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation". In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000.

[57] FRANCIS, W. N., KUČERA, H. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Relatório técnico, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1964.

[58] HINRICHS, E. W., BARTELS, J., KAWATA, Y., et al. "The VERBMOBIL Treebanks". In: *KONVENS 2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation"*, p. 107–112. VDE-Verlag GmbH, 2000.

[59] AFONSO, S., BICK, E., HABER, R., et al. ""Floresta Sintá(c)tica": a treebank for Portuguese". In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation LREC-2002*, p. 1698–1703, 2002.

[60] BOSCO, C., LOMBARDO, V., VASSALLO, D., et al. "Building a Treebank for Italian: a Data-driven Annotation Schema". In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation LREC-2000*, p. 99–105, 2000.

[61] SKUT, W., KRENN, B., BRANTS, T., et al. "An annotation scheme for free word order languages". In: *Proceedings of the 5th conference on Applied natural language processing*, ANLC '97, p. 88–95. Association for Computational Linguistics, 1997.

[62] BOGUSLAVSKY, I., CHARDIN, I., GRIGORIEVA, S., et al. "Development of a Dependency Treebank for Russian and its possible Applications in NLP". In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation LREC-2002*, p. 852–856, 2002.

[63] OFLAZER, K., SAY, B., HAKKANI-TÜR, D. Z., et al. "Building a Turkish treebank", *Treebanks*, p. 261–277, 2003.

[64] HEATON, J. *Programming Neural Networks with Encog 2 in Java.* Chesterfield, MO, USA, Heaton Research, Inc., 2010.

[65] MCCARTHY, J. J. "A Prosodic Theory of Nonconcatenative Morphology", *Linguistic Inquiry*, v. 12, n. 3, pp. 373–418, 1981.

[66] KIRAZ, G. A. "Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study on Syriac and Arabic", *Computational Linguistics*, v. 26, n. 1, pp. 77–105, mar. 2000.

[67] KIRAZ, G. A. *Computational Nonlinear Morphology: Survey of Semitic Computational Morphology*. New York, NY, USA, Cambridge University Press, 2001.

[68] COHEN-SYGAL, Y., WINTNER, S. "Finite-State Registered Automata for Non-Concatenative Morphology", *Computational Linguistics*, v. 32, n. 1, pp. 49–82, mar. 2006.

[69] DENERO, J., USZKOREIT, J. "Inducing Sentence Structure from Parallel Corpora for Reordering". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 193–203, Edinburgh, Scotland, UK., jul. 2011. Association for Computational Linguistics.

[70] DE LA HIGUERA, C. *Grammatical Inference: Learning Automata and Grammars*. New York, NY, USA, Cambridge University Press, 2010.

Apêndice A

Adquirindo o índice de síntese

Neste apêndice serão mostrados os dez primeiros artigos da Declaração Universal de Direitos Humanos, devidamente divididos morfemicamente. Os morfemas derivacionais serão expostos em **negrito**, os inflexionais em *itálico* e as raízes (morfemas composicionais) em fonte regular. Ao fim da seção de cada língua, o valor do seu índice de síntese é calculado, assim como os outros índices menores. Estes valores são representados como intervalos com 95% de confiança. A divisão morfêmica destes artigos foi feita por somente uma pessoa para fins de análise do comportamento e do desempenho do etiquetador.

A.1 Mandarim

1. Rén rén shēng ér zì-yóu, zài zūn-yán hé quán-lì shàng yī-lǜ píng-děng. Tā-men fù-yǒu lǐ-xìng hé liáng-xīn, bìng yīng yǐ xiōng-dì guān-xì de jīng-shén xiāng duì-dài.

2. Rén rén yǒu zì-gé xiǎng-yǒu běn xuān-yán suǒ zài de yī-qiè quán-lì hé zì-yóu, bù fēn zhǒng-zú, fū-sè, xìng-bié, yǔ-yán, zōng-jiào, zhèng-zhì huò qí-tā jiàn-jiě, guó-jí huò shè-huì chū-shēn, cái-chǎn, chū-shēng huò qí-tā shēn-fèn děng rèn-hé qū-bié.

   Bìng-qiě bù-dé yīn yī-rén suǒ-shǔ de guó-jiā huò lǐng-tǔ de zhèng-zhì de, xíng-zhèng de huò-zhě guó-jì dì dì-wèi zhī bù-tóng ér yǒu suǒ qū-bié, wú-lùn gāi lǐng-tǔ shì dú-lì lǐng-tǔ, tuō-guǎn lǐng-tǔ, fēi zì-zhì lǐng-tǔ huò-zhě chǔ-yú qí-tā rèn-hé zhǔ-quán shòu xiàn-zhì de qíng-kuàng zhī xià.

3. Rén rén yǒu quán xiǎng-yǒu shēng-mìng, zì-yóu hé rén-shēn ān-quán.

4. Rèn-hé rén bù-dé shǐ wèi nú-lì huò nú-yì; yī qiē xíng-shì de nú-lì zhì-dù hé nú-lì mǎi-mài, jūn yīng yǔ-yǐ jìn-zhǐ.

5. Rèn-hé rén bù-dé jiā-yǐ kù-xíng, huò shī yǐ cán-rěn de, bù-rén-dào de huò wǔ-rǔ xìng de dài-yù huò xíng-fá.

6. Rén rén zài rèn-hé dì-fāng yǒu quán bèi chéng-rèn zài fǎ-lǜ qián de rén-gé.

7. Fǎ-lǜ zhī-qián rén rén píng-děng, bìng yǒu quán xiǎng-shòu fǎ-lǜ de píng-děng bǎo-hù, bù shòu rèn-hé qí-shì. Rén rén yǒu quán xiǎng-shòu píng-děng bǎo-hù, yǐ-miǎn shòu wéi-fǎn běn xuān-yán de rèn-hé qí-shì xíng-wéi yǐ-jí shān-dòng zhè zhǒng qí-shì de rèn-hé xíng-wéi zhī hài.

8. Rèn-hé rén dāng xiàn-fǎ huò fǎ-lǜ suǒ fù-yǔ tā de jī-běn quán-lì zāo-shòu qīn-hài shí, yǒu quán yóu hé-gé de guó-jiā fǎ-tíng duì zhè zhǒng qīn-hài xíng-wéi zuò yǒu-xiào de bǔ-jiù.

9. Rèn-hé rén bù-dé jiā-yǐ rèn-yì dài-bǔ, jū-jìn huò fàng-zhú.

10. Rén rén wán-quán píng-děng dì yǒu quán yóu yī-gè dú-lì ér wú piān-yǐ de fǎ-tíng jìn-xíng gōng-zhèng de hé gōng-kāi de shěn-xùn, yǐ què-dìng tā de quán-lì hé yì-wù bìng pàn-dìng duì tā tí-chū de rèn-hé xíng-shì zhǐ-kòng.

$CIS = \frac{452}{294} \approx 1,5374 \pm 0,0579$ (137 palavras com uma raiz, 156 com duas e 1 com três);
$DIS = \frac{0}{294} = 0$ (não há morfema derivacional);
$IIS = \frac{0}{294} = 0$ (não há morfema inflexional);
$IS = \frac{452}{294} \approx 1,5374 \pm 0,0579$ (137 palavras com um morfema, 156 com dois e 1 com três).

A.2 Inglês

1. All human being-*s* are born free and equal in dign-**ity** and right-*s*. They are endow-*ed* with reason and consci-**ence** and should act to-**ward**-*s* one an-other in a spirit of brother-**hood**.

2. Every-one is **en-**title-*d* to all the right-*s* and free-**dom**-*s* set forth in this Declarat-**ion**, with-out distinct-**ion** of any kind, such as race, colour, sex, langu-**age**, religion, polit-**ic-al** or other opinion, nation-**al** or soci-**al** origin, proper-**ty**, birth or other status.

   Furth-*er*-more, no distinct-**ion** shall be ma-*de* on the bas-**is** of the polit-**ic-al**, juris-dict-**ion-al** or **inter-**nation-**al** status of the country or territory to which a person belong-*s*, whether it be **in-**depend-**ent**, trust, **non-**self-govern-*ing* or under any other limitat-**ion** of sovereign-**ty**.

3. Every-one ha-*s* the right to life, liber-**ty** and secur-**ity** of person.

4. No one shall be held in slave-**ry** or serv-**itude**; slave-**ry** and the slave trade shall be prohibit-*ed* in all their form-*s*.

5. No one shall be subject-*ed* to torture or to cruel, **in**-human or **de**-grad-*ing* treat-**ment** or punish-**ment**.

6. Every-one ha-*s* the right to **re**-cognit-**ion** every-where as a person **be**-fore the law.

7. All are equal **be**-fore the law and are **en**-titl-*ed* with-out any discriminat-**ion** to equal protect-**ion** of the law. All are **en**-titl-*ed* to equal protect-**ion** against any discriminat-**ion** in violat-**ion** of this Declarat-**ion** and against any incite-**ment** to such discriminat-**ion**.

8. Every-one ha-*s* the right to an effect-**ive** remedy by the compet-**ent** nation-**al** tribunal-*s* for act-*s* violat-*ing* the funda-**ment-al** right-*s* grant-*ed* him by the constitut-**ion** or by law.

9. No one shall be subject-*ed* to arbitr-**ary** arrest, detent-**ion** or exile.

10. Every-one is **en**-titl-*ed* in full equal-**ity** to a fair and publ-**ic** hear-*ing* by an **in**-depend-**ent** and **im**-parti-**al** tribunal, in the determinat-**ion** of his right-*s* and obligat-**ion**-*s* and of any crimin-**al** charge against him.

$CIS = \frac{291}{279} \approx 1,0430 \pm 0,0238$ (267 palavras com uma raiz e 12 com duas);
$DIS = \frac{68}{279} \approx 0,2437 \pm 0,0586$ (220 palavras sem morfema derivacional, 50 com um e 9 com dois);
$IIS = \frac{30}{279} \approx 0,1075 \pm 0,0364$ (249 palavras sem morfema inflexional e 30 com um);
$IS = \frac{291 + 68 + 30}{279} \approx 1,3943 \pm 0,0747$ (190 palavras com um morfema, 70 com dois, 17 com três e 2 com quatro).

A.3 Japonês

1. Sube-*te* no nin-gen wa, umare-*nagara* ni shite ji-yū de ar-*i*, katsu, son-gen to ken-ri to ni tsui-*te* byō-dō de ar-*u*. Nin-gen wa, ri-sei to ryō-shin to o sazuke-*rare-te* o-*ri*, tagai-**ni** dō-hō no sei-shin o mot-*te* kō-dō shi-*nak-ereba* nar-*anai*.

2. Sube-*te* hito wa, jin-shu, hi-fu no iro, sei, gen-go, shū-kyō, sei-ji-**jō** so-**no**-hoka no i-ken, koku-min-**teki** moshiku-**wa** sha-kai-**teki** shus-shin, zai-san, mon-chi so-**no**-hoka no chi-i mata-**wa** ko-**re** ni rui s-*uru* ika-**nar**-*u* ji-yū ni yor-*u* sa-betsu o mo uke-*ru* koto nak-*u*, ko-**no** sen-gen ni kakage-*ru* sube-*te* no ken-ri to ji-yū to o kyō-yū s-*uru* koto ga de-k-*iru*.

   Sara-**ni**, ko-jin no zoku s-*uru* kuni mata-**wa** chi-iki ga doku-ritsu-koku de ar-*u* to, shin-taku-tō-chi chi-iki de ar-*u* to, hi-ji-chi chi-iki de ar-*u* to, mata-**wa** hoka no nan-ra-**ka** no shu-ken sei-gen no shita ni ar-*u* to o to-**wazu**, so-**no** kuni mata-**wa** chi-iki no sei-ji-**jō**, kan-katsu-**jō** mata-**wa** koku-sai-**jō** no chi-i ni moto-zuk-*u* ika-**nar**-*u* sa-betsu mosh-*ite* wa nar-*anai*.

3. Sube-*te* no hito wa, sei-mei, ji-yū oyob-**i** shin-tai no an-zen ni tai s-*uru* ken-ri o yū s-*uru*.

4. Nan-pito mo, do-rei ni s-*are*, mata-**wa** ku-eki ni fuku s-*uru* koto wa nai. Do-rei sei-do oyob-**i** do-rei bai-bai wa, ika-**nar**-*u* katachi ni oi-*te* mo kin-shi s-*uru*.

5. Nan-pito mo, gō-mon mata-**wa** zan-gyaku na, hi-jin-dō-**teki** na moshiku-**wa** kutsu-joku-**teki** na tori-atsukai moshiku-**wa** kei-batsu o uke-*ru* koto wa nai.

6. Sube-*te* no hito wa, ika-**nar**-*u* ba-sho ni oi-*te* mo, hō no shita ni oi-*te*, hito to-sh-*ite* mitome-*rare-ru* ken-ri o yū s-*uru*.

7. Sube-*te* no hito wa, hō no shita ni oi-*te* byō-dō de ar-*i*, mata, ika-**nar**-*u* sa-betsu mo nashi ni hō no byō-dō na ho-go o uke-*ru* ken-ri o yū s-*uru*. Sube-*te* no hito wa, ko-**no** sen-gen ni i-han s-*uru* ika-**nar**-*u* sa-betsu ni tai sh-*ite* mo, mata, so-**no** yō na sa-betsu o sosonokas-*u* ika-**nar**-*u* kō-i ni tai sh-*ite* mo, byō-dō na ho-go o uke-*ru* ken-ri o yū s-*uru*.

8. Sube-*te* no hito wa, ken-pō mata-**wa** hō-ritsu ni yot-*te* atae-*rare-ta* ki-hon-**teki** ken-ri o shin-gai s-*uru* kō-i ni tai sh-*i*, ken-gen o yū s-*uru* koku-nai sai-ban-**sho** ni yor-*u* kōka-**teki** na kyū-sai o uke-*ru* ken-ri o yū s-*uru*.

9. Nan-pito mo, hoshīmama ni tai-ho, kō-kin, mata-**wa** tsui-hō s-*are-ru* koto wa nai.

10. Sube-*te* no hito wa, ji-ko no ken-ri oyob-**i** gi-mu narab-**i-ni** ji-ko ni tai s-*uru* kei-ji seki-nin ga ket-tei s-*are-ru* ni atat-*te*, doku-ritsu no kō-hei na sai-ban-**sho** ni yor-*u* kō-hei na kō-kai no shin-ri o uke-*ru* koto ni tsui-*te* kan-zen ni byō-dō no ken-ri o yū s-*uru*.

$CIS = \frac{511}{390} \approx 1,3103 \pm 0,0492$ (274 palavras com uma raiz, 112 com duas, 3 com três e 1 com quatro);

$DIS = \frac{47}{390} \approx 0,1205 \pm 0,0331$ (344 palavras sem morfema derivacional, 45 com um e 1 com dois);

$IIS = \frac{84}{390} \approx 0,2154 \pm 0,0444$ (312 palavras sem morfema inflexional, 72 com um e 6 com dois);

$IS = \frac{511 + 47 + 84}{390} \approx 1,6462 \pm 0,0647$ (174 palavras com um morfema, 182 com dois, 32 com três e 2 com quatro).

A.4 Português

1. Tod-*o-s* o-*s* ser-*es* human-*o-s* nasc-*e-m* livre-*s* e iguai-*s* em dign-**idade** e direito-*s*. S-*ão* dot-*a-d-o-s* de razão e **con**-sci-**ência** e dev-*e-m* ag-*i-r* em rela-**ção** un-*s* a-o-*s* outr-*o-s* com espírito de fratern-**idade**.

2. Tod-*o* ser human-*o* t-*e-m* capac-**idade** para goz-*a-r* o-*s* direito-*s* e a-*s* liber-**dade**-*s* estabelec-*i-d-o-s* n-est-*a* Declara-**ção**, sem distin-**ção** de qual-quer espécie, sej-*a* de raça, cor, sexo, idioma, religião, opinião polít-**ic**-*a* ou de outr-*a* natur-**eza**, origem nacion-**al** ou soci-**al**, riqu-**eza**, nasci-**mento**, ou qual-quer outr-*a* condi-**ção**.

   Não s-*e-r-á* tam-bém feit-*a* nen-hum-*a* distin-**ção** fund-*a-d-a* n-a condi-**ção** polít-**ic**-*a*, juríd-**ic**-*a* ou **inter**-nacion-**al** d-o país ou território a que pertenç-*a* um-*a* pessoa, quer se trat-*e* de um território **in**-depend-*e-nte*, sob tutela, sem governo própri-*o*, quer sujeit-*o* a qual-quer outr-*a* limita-**ção** de soberan-**ia**.

3. Tod-*o* ser human-*o* t-*e-m* direito a-a vida, a-a liber-**dade** e a-a segur-**ança** pesso-**al**.

4. Nin-**guém** s-*e-r-á* mant-*i-d-o* em escrav-**idão** ou serv-**idão**; a escrav-**idão** e o tráfic-*o* de escrav-*o-s* s-*e-r-ão* proib-*i-d-o-s* em tod-*a-s* a-*s* su-*a-s* form-*a-s*.

5. Nin-**guém** s-*e-r-á* **sub**-met-*i-d-o* a-a tortur-*a* nem a trata-**mento** ou castig-*o* cruel, **des**-uman-*o* ou **de**-grad-*a-nte*.

6. Tod-*o* ser human-*o* t-*e-m* o direito de s-*e-r*, em tod-*o-s* o-*s* lugar-*es*, **re**-conhec-*i-d-o* como pessoa per-ante a lei.

7. Tod-*o-s* s-*ão* iguai-*s* per-ante a lei e t-*ê-m* direito, sem qual-quer distin-**ção**, a igual prote-**ção** d-a lei. Tod-*o-s* t-*ê-m* direito a igual prote-**ção** contra qual-quer discrimina-**ção** que viol-*e* a presente Declara-**ção** e contra qual-quer incita-**mento** a tal discrimina-**ção**.

8. Tod-*o* ser human-*o* t-*e-m* direito a receb-*e-r* d-o-*s* tribunai-*s* nacion-**ai**-*s* compet-*e-nte-s* remédi-*o* efet-**iv**-*o* para o-*s* ato-*s* que viol-*e-m* o-*s* direito-*s*

funda-**ment-ai**-*s* que lhe sej-*a-m* **re-**conhec-*i-d-o-s* pel-a constitui-**ção** ou pel-a lei.

9. Nin-**guém** s-*e-r-á* arbitr-**aria-mente** pres-*o*, det-*i-d-o* ou exil-*a-d-o*.

10. Tod-*o* ser human-*o* t-*e-m* direito, em plen-*a* igual-**dade**, a um-*a* just-*a* e públ-**ic**-*a* audi-**ência** por parte de um tribunal **in-**depend-*e-nte* e **im-**parci-**al**, para decid-*i-r* sobre seu-*s* direito-*s* e dever-*es* ou d-o funda-**ment**-*o* de qual-quer acusa-**ção** crimin-**al** contra el-*e*.

$CIS = \frac{318}{294} \approx 1,0816 \pm 0,0314$ (270 palavras com uma raiz e 24 com duas);
$DIS = \frac{63}{294} \approx 0,2143 \pm 0,0515$ (236 palavras sem morfema derivacional, 53 com um e 5 com dois);
$IIS = \frac{177}{294} \approx 0,6020 \pm 0,1053$ (182 palavras sem morfema inflexional, 67 com um, 29 com dois, 12 com três e 4 com quatro);
$IS = \frac{318+63+177}{294} \approx 1,8980 \pm 0,1111$ (120 palavras com um morfema, 111 com dois, 43 com três, 14 com quatro, 5 com cinco e 1 com seis).

A.5 Italiano

1. Tutt-*i* gl-*i* esser-*i* uman-*i* nasc-*o-no* liber-*i* ed egual-*i* in dign-**ità** e diritt-*i*. Ess-*i* s-*o-no* dot-*a-t-i* di ragion-*e* e di **co-**sci-**enz**-*a* e dev-*o-no* ag-*i-re* gl-*i* un-*i* verso gl-*i* altr-*i* in spirit-*o* di fratell-**anz**-*a*.

2. Ad ogn-*i* individu-*o* spett-*a-no* tutt-*i* i diritt-*i* e tutt-*e* l-*e* liber-**tà** enunci-*a-t-e* n-ell-*a* present-*e* Dichiara-**zion**-*e*, senza distin-**zion**-*e* **alc-**un-*a*, per ragion-*i* di razz-*a*, di color-*e*, di sess-*o*, di lingu-*a*, di religion-*e*, di opinion-*e* polit-**ic**-*a* o di altr-*o* gener-*e*, di origin-*e* nazion-**al**-*e* o soci-**al**-*e*, di ricch-**ezz**-*a*, di nasc-**it**-*a* o di altr-*a* condi-**zion**-*e*.

   Ness-un-*a* distin-**zion**-*e* s-*a-r-à* in-oltr-*e* stabil-*i-t-a* su-ll-*a* bas-*e* de-ll-*o* statut-*o* polit-**ic**-*o*, giurid-**ic**-*o* o **inter-**nazion-**al**-*e* de-l paes-*e* o de-l territori-*o* cui un-*a* person-*a* appart-*ien-e*, s-*ia* **in-**dipend-*e-nt-e*, o **sotto-**post-*o* ad amministra-**zion**-*e* fiduci-**ari**-*a* o non auto-nom-*o*, o soggett-*o* a qual-s-*ia-si* limita-**zion**-*e* di sovran-**ità**.

3. Ogn-*i* individu-*o* ha diritt-*o* a-ll-*a* vit-*a*, a-ll-*a* liber-**tà** ed a-ll-*a* sicur-**ezz**-*a* de-ll-*a* propri-*a* person-*a*.

4. **Ness-**un individu-*o* pot-*r-à* ess-*e-re* ten-*u-t-o* in stat-*o* di schiav-**itù** o di serv-**itù**; l-*a* schiav-**itù** e l-*a* tratt-*a* de-gl-*i* schiav-*i* s-*a-r-a-nno* proib-*i-t-e* sotto qual-s-*ia-si* form-*a*.

5. **Ness-**un individu-*o* pot-*r*-*à* ess-*e*-*re* **sotto-**post-*o* a tortur-*a* o a tratta-**ment**-*o* o a puni-**zion**-*e* crudel-*i*, **in-**uman-*i* o **de-**grad-*a*-*nt*-*i*.

6. Ogn-*i* individu-*o* ha diritt-*o*, in ogn-*i* luog-*o*, a-l **ri-**conosci-**ment**-*o* de-ll-*a* su-*a* person-**al-ità** giurid-**ic**-*a*.

7. Tutt-*i* s-*o*-*no* egual-*i* dinanzi a-ll-*a* legg-*e* e h-*a*-*nno* diritt-*o*, senza **alc-**un-*a* discrimina-**zion**-*e*, ad un-*a* egual-*e* tutel-*a* da part-*e* de-ll-*a* legg-*e*. Tutt-*i* h-*a*-*nno* diritt-*o* ad un-*a* egual-*e* tutel-*a* contro ogn-*i* discrimina-**zion**-*e* che viol-*i* l-*a* present-*e* Dichiara-**zion**-*e* come contro qual-s-*ia*-*si* incita-**ment**-*o* a tal-*e* discrimina-**zion**-*e*.

8. Ogn-*i* individu-*o* ha diritt-*o* ad un' effett-**iv**-*a* poss-**ibil-ità** di ricors-*o* a compet-*e*-*nt*-*i* tribunal-*i* contro att-*i* che viol-*i*-*no* i diritt-*i* fonda-**ment-al**-*i* a lui **ri-**conosci-*u*-*t*-*i* da-ll-*a* costitu-**zion**-*e* o da-ll-*a* legg-*e*.

9. **Ness-**un individu-*o* pot-*r*-*à* ess-*e*-*re* arbitr-**aria-mente** arrest-*a*-*t*-*o*, deten-*u*-*t*-*o* o esili-*a*-*t*-*o*.

10. Ogn-*i* individu-*o* ha diritt-*o*, in posi-**zion**-*e* di pien-*a* uguagli-**anz**-*a*, ad un-*a* equ-*a* e pubbl-**ic**-*a* udi-**enz**-*a* davanti ad un tribunal-*e* **in-**dipend-*e*-*nt*-*e* e **im-**parzi-**al**-*e*, a-l fin-*e* de-ll-*a* determina-**zion**-*e* de-i su-*o*-*i* diritt-*i* e de-i su-*o*-*i* dover-*i*, non-ché de-ll-*a* fondat-**ezz**-*a* di ogn-*i* accus-*a* pen-**al**-*e* che gl-*i* veng-*a* rivolt-*a*.

$CIS = \frac{339}{312} \approx 1,0865 \pm 0,0312$ (285 palavras com uma raiz e 27 com duas);
$DIS = \frac{69}{312} \approx 0,2211 \pm 0,0525$ (251 palavras sem morfema derivacional, 53 com um e 8 com dois);
$IIS = \frac{254}{312} \approx 0,8141 \pm 0,0856$ (110 palavras sem morfema inflexional, 166 com um, 21 com dois, 14 com três e 1 com quatro);
$IS = \frac{339+69+254}{312} \approx 2,1218 \pm 0,1052$ (90 palavras com um morfema, 122 com dois, 77 com três, 18 com quatro e 5 com cinco).

A.6 Alemão

1. All-*e* Mensch-*en* sind frei und gleich an Würd-*e* und Recht-*en* *ge*-bor-*en*. Sie sind mit Vernunft und **Ge-**wissen **be-**gab-*t* und soll-*en* ein-ander i-*m* Geist d-*er* Brüder-**lich-keit** **be-**gegn-*en*.

2. Jed-*er* ha-*t* **An-**spruch auf d-*ie* in dies-*er* **Er-**klär-**ung** **ver-**künd-*et*-*en* Recht-*e* und Frei-**heit**-*en* ohne irgend-ein-*en* **Unter-**schied, etwa nach Rasse,

Haut-farbe, Geschlecht, Sprach-*e*, Religion, polit-**isch**-*er* oder sonst-**ig**-*er* **Über**-zeug-**ung**, nation-**al**-*er* oder sozi-**al**-*er* **Her**-kunft, **Ver**-mögen, Geburt oder sonst-**ig**-*em* Stand.

D-*es* weiter-*en* darf **k**-ein **Unter**-schied *ge*-mach-*t* werd-*en* auf Grund d-*er* polit-**isch**-*en*, recht-**lich**-*en* oder **inter**-nation-**al**-*en* Stell-**ung** d-*es* Land-*es* oder Gebiet-*s*, d-*em* ein-*e* Person **an**-*ge*-hör-*t*, gleich-gült-**ig**, ob dies-*es* **un-ab**-häng-**ig** ist, unter Treu-hand-**schaft** steh-*t*, **k**-ein-*e* Selbst-reg-**ier-ung** **be**-sitz-*t* oder sonst in sein-*er* Souverän-**ität** **ein**-*ge*-schränk-*t* ist.

3. Jed-*er* ha-*t* da-*s* Recht auf Leben, Frei-**heit** und Sicher-**heit** d-*er* Person.

4. Nie-mand darf in Sklav-**erei** oder Leib-eigen-**schaft** *ge*-halt-*en* werd-*en*; Sklav-**erei** und Sklav-*en*-handel sind in all-*en* ih-*r*-*en* Form-*en* **ver**-bot-*en*.

5. Nie-mand darf d-*er* Folt-*er* oder grau-**sam**-*er*, **un**-mensch-**lich**-*er* oder **er**-niedr-**ig**-*end*-*er* **Be**-handl-**ung** oder Straf-*e* **unter**-worf-*en* werd-*en*.

6. Jed-*er* ha-*t* da-*s* Recht, über-all als recht-*s*-fähig **an-er**-kann-*t* zu werd-*en*.

7. All-*e* Mensch-*en* sind vor d-*em* **Ge**-setz gleich und hab-*en* ohne **Unter**-schied **An**-spruch auf gleich-*en* Schutz durch da-*s* **Ge**-setz. All-*e* hab-*en* **An**-spruch auf gleich-*en* Schutz gegen jed-*e* Diskrimin-**ier-ung**, d-*ie* gegen dies-*e* **Er**-klär-**ung** **ver**-stöß-*t*, und gegen jed-*e* **Auf**-hetz-**ung** zu ein-*er* derar-**tig**-*en* Diskrimin-**ier-ung**.

8. Jed-*er* ha-*t* **An**-spruch auf ein-*en* wirk-**sam**-*en* Recht-*s*-**be**-helf bei d-*en* **zu**-ständ-**ig**-*en* inner-staat-**lich**-*en* **Ge**-richt-*en* gegen Handl-**ung**-*en*, durch d-*ie* sein-*e* ih-*m* nach d-*er* **Ver**-fass-**ung** oder nach d-*em* **Ge**-setz **zu**-steh-*end*-*en* Grund-recht-*e* **ver**-letz-*t* werd-*en*.

9. Nie-mand darf will-kür-**lich** fest-*ge*-nomm-*en*, in Haft *ge*-halt-*en* oder d-*es* Land-*es* **ver**-wies-*en* werd-*en*.

10. Jed-*er* ha-*t* bei d-*er* Fest-stell-**ung** sein-*er* Recht-*e* und Pflicht-*en* so-wie bei ein-*er* gegen ih-*n* **er**-hob-*en*-*en* straf-recht-**lich**-*en* **Be**-schuld-**ig-ung** in voll-*er* Gleich-**heit** **An**-spruch auf ein gerecht-*es* und öffen-*t*-**lich**-*es* **Ver**-fahr-*en* vor ein-*em* **un-ab**-häng-**ig**-*en* und **un**-partei-**isch**-*en* **Ge**-richt.

$CIS = \frac{289}{268} \approx 1,0784 \pm 0,0322$ (247 palavras com uma raiz e 21 com duas);

$DIS = \frac{101}{268} \approx 0,3769 \pm 0,0779$ (189 palavras sem morfema derivacional, 60 com um, 16 com dois e 3 com três);

$IIS = \frac{140}{268} \approx 0,5224 \pm 0,0706$ (141 palavras sem morfema inflexional, 114 com um e 13 com dois);

$IS = \frac{289 + 101 + 140}{268} \approx 1,9776 \pm 0,1111$ (94 palavras com um morfema, 108 com dois, 46 com três, 18 com quatro e 2 com cinco).

A.7 Russo

1. Vs-*e* ljud-*i* rožd-*a-jut-sja* svobod-**n**-*ymi* i rav-**n**-*ymi* v svo-*em* dost-**oin-stv**-*e* i prav-*ah*. On-*i* nadel-*e-n-y* razum-*om* i sovest-*'ju* i dolžn-*y* postup-*a-t'* v otnoš-**eni**-*i* drug drug-*a* v duh-*e* brat-**stv**-*a*.

2. Každ-*yj* čelovek dolžen oblad-*a-t'* vs-*emi* prav-*ami* i vs-*emi* svobod-*ami*, provoz-glaš-**enn**-*ymi* nasto-*ja-šč-ej* Deklaraci-*ej*, bez kak-*ogo* by to ni b-*yl-o* različ-**i**-*ja*, kak-to v otnoš-**eni**-*i* ras-*y*, cvet-*a* kož-*i*, pol-*a*, jazyk-*a*, religi-*i*, političe-**esk**-*ih* ili in-*yh* ubežd-**eni**-*j*, nacional'-**n**-*ogo* ili social'-**n**-*ogo* proish-**ožd-eni**-*ja*, imušč-**estv-enn**-*ogo*, sosl-**ov-n**-*ogo* ili in-*ogo* polož-**eni**-*ja*.

   Krome t-*ogo*, ne dolžn-*o* provod-*i-t'-sja* ni-kak-*ogo* različ-**i**-*ja* na osnov-*e* političe-**esk**-*ogo*, prav-**ov**-*ogo* ili meždu-narodn-*ogo* status-*a* stran-*y* ili territori-*i*, k kotor-*oj* čelovek **pri**-nadlež-*it*, **ne**-zavis-**im**-*o* ot t-*ogo*, javl-*ja-et-sja* li èt-**a** territori-*ja* **ne**-zavis-**im**-*oj*, **pod**-opeč-**n**-*oj*, **ne-samo**-upravl-*ja-jušč-ej-sja* ili kak-libo inače ograniče-**enn**-*oj* v svo-*em* suveren-**itet**-*e*.

3. Každ-*yj* čelovek im-*e-et* prav-*o* na žizn-*'*, na svobod-*u* i na ličn-*uju* **ne-pri**-kosn-**ov-enn-ost**-*'*.

4. Ni-kto ne dolžen soderž-*a-t'-sja* v rab-**stv**-*e* ili v **pod**-nevol'-**n**-*om* sosto-*ja-ni-i*; rab-**stv**-*o* i rab-*o*-torg-**ov-l**-*ja* zapr-**eščč**-*a-jut-sja* vo vs-*eh* ih vid-*ah*.

5. Ni-kto ne dolžen **pod**-verg-*a-t'-sja* pyt-**k**-*am* ili žestok-*im*, **bes**-čelovеč-**n**-*ym* ili uniž-*a-jušč-im* ego dost-**oin-stv**-*o* obraščč-**eni**-*ju* i nakaz-**a-ni**-*ju*.

6. Každ-*yj* čelovek, gde by on ni nahod-*i-l-sja*, im-*e-et* prav-*o* na prizn-**a-ni**-*e* ego prav-*o*-sub"ekt-**n-ost**-*i*.

7. Vs-*e* ljud-*i* ravn-*y* pered zakon-*om* i im-*e-jut* prav-*o*, bez vsjak-*ogo* različ-**i**-*ja*, na ravn-*uju* zaščit-*u* zakon-*a*. Vs-*e* ljud-*i* im-*e-jut* prav-*o* na ravn-*uju* zaščit-*u* ot kak-*oj* by to ni b-*yl-o* diskriminaci-*i*, naruš-*a-jušč-ej* nasto-*ja-šč-uju* Deklaraci-*ju*, i ot kak-*ogo* by to ni b-*yl-o* **pod**-strek-**a-tel**-*'*-**stv**-*a* k tak-*oj* diskriminaci-*i*.

8. Každ-*yj* čelovek im-*e*-*et* prav-*o* na èffektiv-**n**-*oe* **vos**-stan-**ov**-**l**-**eni**-*e* v prav-*ah* kompetent-**n**-*ymi* nacional'-**n**-*ymi* sud-*ami* v sluča-*jah* naruš-**eni**-*ja* ego osnov-**n**-*yh* prav, **pre**-dost-**av**-**l**-*enn*-*yh* emu konstituci-*ej* ili zakon-*om*.

9. Ni-kto ne mož-*et* b-*y*-*t'* **pod**-vergn-*ut* proiz-**vol'**-**n**-*omu* arest-*u*, zaderž-**a**-**ni**-*ju* ili izgn-**a**-**ni**-*ju*.

10. Každ-*yj* čelovek, dlja **o**-predel-**eni**-*ja* ego prav i objaz-**a**-**nn**-**ost**-*ej* i dlja usta-**nov**-**l**-**eni**-*ja* **ob**-osnov-**a**-**nn**-**ost**-*i* pred"-javl-**enn**-*ogo* emu ugol-**ov**-**n**-*ogo* obvin-**eni**-*ja*, im-*e*-*et* prav-*o*, na osnov-*e* poln-*ogo* raven-**stv**-*a*, na to, čto-by ego delo b-*yl*-*o* **ras**-smotr-**en**-*o* glas-**n**-*o* i s sobljud-**eni**-*em* vs-*eh* trebov-**a**-**ni**-*j* spravedliv-**ost**-*i* **ne**-zavis-**im**-*ym* i **bes**-**pri**-strast-**n**-*ym* sud-*om*.

$CIS = \frac{291}{279} \approx 1,0430 \pm 0,0238$ (267 palavras com uma raiz e 12 com duas);
$DIS = \frac{113}{279} \approx 0,4050 \pm 0,0978$ (210 palavras sem morfema derivacional, 38 com um, 23 com dois, 4 com três, 3 com quatro e 1 com cinco);
$IIS = \frac{220}{279} \approx 0,7885 \pm 0,0919$ (105 palavras sem morfema inflexional, 143 com um, 17 com dois, 13 com três e 1 com quatro);
$IS = \frac{291 + 113 + 220}{279} \approx 2,2366 \pm 0,1513$ (98 palavras com um morfema, 87 com dois, 48 com três, 33 com quatro, 5 com cinco, 5 com seis e 3 com sete).

A.8 Turco

1. Bütün insan-*lar* hür, hays-**iyet** ve hak-*lar* bak-**ım**-*ın*-*dan* eş-**it** doğ-*ar*-*lar*. Akıl ve vicdan-*a* sahip-*tir*-*ler* ve bir-bir-*ler*-*in*-*e* karşı kardeş-**lik** zihn-**iyet**-*i* ile hareket et-*meli*-*dir*-*ler*.

2. Her-**kes**, ırk, renk, cins-**iyet**, dil, din, siyas-**i** ve-ya diğer her-hangi bir akide, mill-**i** ve-ya içtima-**i** menşe, servet, doğ-**uş** ve-ya her-hangi diğer bir fark göz-**e**-*t*-*il*-*mek*-*sizin* iş-bu Beyan-name-*de* ilan ol-*un*-*an* tekmil hak-*lar*-*dan* ve bütün hürr-**iyet**-*ler*-*den* istifade ed-*ebil*-*ir*.

   Bun-*dan* başka, bağ-**ım**-*sız* memleket uyruğ-*u* ol-*sun*, vesa-**yet** alt-*ın*-*da* bulun-*an*, gayri muhtar ve-ya sair bir egemen-**lik** kayıt-**la**-*ma*-*sın*-*a* tabi ülke uyruğ-*u* ol-*sun*, bir şahıs hakk-*ın*-*da*, uyruğ-*u* bulun-*duğ*-*u* memleket ve-ya ülke-*nin* siyas-**i**, hukuk-**i** ve-ya millet-*ler*-ara-**sı** statü-*sü* bak-**ım**-*ın*-*dan* hiç-bir ayrı-**lık** göz-**e**-*t*-*il*-*me*-*yecek*-*tir*.

3. Yaşa-*mak*, hürr-**iyet** ve kişi emn-**iyet**-*i* her ferd-*in* hakk-*ı*-*dır*.

4. Hiç kim-se köle-**lik** ve-ya kul-**luk** alt-*ın*-*da* bulun-*dur*-*ul*-*a*-*ma*-*z*; köle-**lik** ve köle ticaret-*i* her türlü şekl-*i*-*yle* yasak-*tır*.

5. Hiç kim-se işkence-*ye*, zalim-**ane**, gayri-insan-**i**, hays-**iyet** kırıcı ceza-*lar-a* ve-ya muamele-*ler-e* tabi tut-*ul-a-ma-z*.

6. Her-**kes** her ne-*re-de* ol-*ur-sa* ol-*sun* hukuk kişi-**liğ**-*i-nin* tanı-*n-ma-sı* hakk-*ın-ı* haiz-*dir*.

7. Kanun ön-*ün-de* her-**kes** eş-**it**-*tir* ve fark-*sız* ol-*arak* kanun-*un* eş-**it** koru-*ma-sın-dan* istifade hakk-*ın-ı* haiz-*dir*. Her-**kes**-*in* iş-bu Beyan-name-*ye* aykırı her türlü ayırd-ed-**ici** muamele-*ye* karşı ve böyle bir ayırd-ed-**ici** muamele için yap-*ıl-acak* her türlü kışkırt-*ma-ya* karşı eş-**it** koru-*n-ma* hakk-*ı* var-*dır*.

8. Her şahs-*ın* kendin-*e* ana-yasa ve-ya kanun ile tanı-*n-an* ana hak-*lar-a* aykırı muamele-*ler-e* karşı fiil-**i** netice ver-*ecek* şekil-*de* mill-**i** mahkeme-*ler-e* müracaat hakk-*ı* var-*dır*.

9. Hiç kim-se keyf-**i** ol-*arak* tut-**uk-la**-*n-a-ma-z*, al-**ı**-ko-*n-ul-ana-ma-z* ve-ya sür-*ül-e-me-z*.

10. Her-**kes**, hak-*lar-ın-ın*, vecibe-*ler-in-in* ve-ya kendi-*sin-e* karşı ceza-**i** mah-**iyet**-*te* her-hangi bir isnad-*ın* tespit-*in-de*, tam bir eş-**it-lik**-*le*, dava-*sın-ın* bağ-**ım**-*sız* ve taraf-*sız* bir mahkeme taraf-*ın-dan* adil bir şekil-*de* ve açık ol-*arak* gör-*ül-me-si* hakk-*ın-a* sahip-*tir*.

$CIS = \frac{265}{236} \approx 1,1229 \pm 0,0420$ (207 palavras com uma raiz e 29 com duas);
$DIS = \frac{52}{236} \approx 0,2203 \pm 0,0555$ (186 palavras sem morfema derivacional, 48 com um e 2 com dois);
$IIS = \frac{172}{236} \approx 0,7288 \pm 0,1369$ (138 palavras sem morfema inflexional, 49 com um, 34 com dois, 8 com três, 4 com quatro e 3 com cinco);
$IS = \frac{265+52+172}{236} \approx 2,0720 \pm 0,1507$ (85 palavras com um morfema, 88 com dois, 42 com três, 12 com quatro, 4 com cinco, 2 com seis, 2 com sete e 1 com oito).