Indução Gramatical para o Português: a Contribuição da Informação Mútua para Descoberta de Relações de Dependência

Diego Pedro Gonçalves da Silva[1], **Thiago Alexandre Salgueiro Pardo**[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

diegopedro@usp.br, taspardo@icmc.usp.br

Resumo. *Indução gramatical é uma tarefa que busca aprender automaticamente estruturas sintáticas a partir de texto. Poucos trabalhos de indução gramatical foram produzidos direcionados para a língua portuguesa. Neste artigo, reproduzidos o trabalho de [Futrell et al. 2019] para a língua portuguesa e o estendemos ao incluir análise de informação mútua para relações sintáticas específicas. Utilizamos dois treebanks anotados e realizamos experimentos utilizando embeddings de dimensões variadas, demonstrando a hipótese de alta informação mútua para palavras em relações de dependência.*

1. Introdução

Na Linguística, sintaxe é definida como o estudo da organização das palavras (em termos de ordenação e estruturação) na formação de sentenças. Esse entendimento é compartilhado por diferentes visões sobre como a sintaxe deve ser formalizada [Chomsky 2014] [Bresnan et al. 2015] . Quase toda aplicação de Processamento de Línguas Naturais (PLN) necessita de algum conhecimento sintático para obter bons resultados, direta ou indiretamente codificados. Revisores gramaticais, sistemas de simplificação de textos e sistemas de extração de informação são algumas das aplicações que se beneficiam da representação explícita da sintaxe. As aplicações baseadas em grandes modelos de língua, por sua vez, acabam adquirindo noções de sintaxe em seu treinamento, mesmo que ela não seja completamente explicitada..

Dada a relevância da sintaxe, a Indução Gramatical (IG), também chamada de *parsing* não supervisionado [Klein e Manning 2004], é uma tarefa de interesse na comunidade de PLN. Apesar de ela ter a finalidade de induzir ("aprender") automaticamente a gramática a partir de dados textuais sem anotações sintáticas [Klein e Manning 2004], vários autores realizam IG como tarefa semi-supervisionada (IGSS) ou supervisionada (IGS) [Headden III et al. 2009] [Spitkovsky et al. 2013]. Sendo assim, utilizaremos o termo (IGNS) para nos referirmos à tarefa de IG não supervisionada. É interessante notar que, independentemente de aplicações computacionais de PLN, a IGNS pode auxiliar em várias frentes. Na Linguística, pode ser útil para aprender a gramática de línguas mortas ou com escassez de recursos (como as indígenas) [Dahl et al. 2023]. Em Psicolinguística, pode ser utilizada para propor modelos de aquisição da linguagem [Bannard et al. 2009]. Em Bioinformática, IGNS é utilizada para inferir estruturas de DNA desconhecidas ou difíceis de serem encontradas em grandes bases de dados [Unold et al. 2020].

A sintaxe (e, por consequência, a IGNS) se vale de duas visões diferentes de representação: a gramática de constituinte e a gramática de dependência. A primeira

estuda como as sentenças são formadas por blocos básicos (sintagmas). No modelo de representação de gramática de dependência, foco desse artigo, estabelecem-se relações de dependências diretamente entre as palavras. A Figura 1 apresenta relações de sujeito (nsubj) (entre o verbo "chutou" e a palavra "menino") e objeto (obj) (entre o verbo e a palavra "bola"), por exemplo.

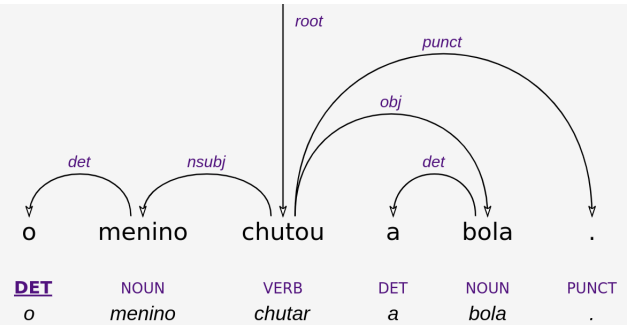


Figura 1. Exemplo de análise de dependência

Devido aos avanços da IGNS nos anos 2000, presumia-se que a IGNS se aproximaria da IGS em desempenho em breve, mas isso ainda não ocorreu [Bod 2007] [Lin et al. 2022]. A maioria dos trabalhos desenvolvidos nas últimas décadas utiliza algum tipo de anotação. Induzir gramática sem nenhuma informação prévia é uma tarefa bastante difícil. Outro desafio é a indução de gramática em sentenças independentemente do tamanho. A maioria dos trabalhos publicados utiliza sentenças de até 10 palavras. Por fim, as diferenças sintáticas entre as línguas apresentam mais um desafio que dificulta a padronização de técnicas para diferentes línguas. Por exemplo, as línguas chinesa, persa e tupi apresentam características linguísticas diferentes por fazerem parte de famílias de línguas diferentes [Theodor e Siebert-Cole 2020], o que pode dificultar a criação de um modelo unificado. No melhor de nosso conhecimento, não encontramos trabalhos publicados de IGNS específico para a língua portuguesa. O trabalho mais similar encontrado foi produzido por [da Costa e Kepler 2014], que implementa uma abordagem semi-supervisionada baseado no trabalho de [Klein e Manning 2004].

Nesse contexto, nosso objetivo neste artigo é explorar a tarefa de IGNS para o português. Em específico, focamos na reprodução de um experimento recente de uso da conhecida medida de Informação Mútua (IM) para tentar predizer palavras que possam estar relacionadas sintaticamente. A IM é uma medida de dependência, assim, quanto maior a informação mútua entre duas palavras, maior a chance de elas estarem relacionadas. Baseamo-nos no trabalho de [Futrell et al. 2019], que, usando IM aplicada a um corpus de milhões de palavras em inglês, mostrou que há uma maior IM entre palavras que mantêm relação de dependência do que entre palavras que não mantêm. Além de avaliar tal técnica para o português, vamos além e verificamos seu comportamento para relações específicas. Realizamos nossos experimentos com dados de *treebanks* alinhados ao modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], amplamente adotado.

Na Seção 2, apresentamos brevemente os principais trabalhos relacionados. Em seguida, na Seção 3, descrevemos a abordagem aplicada no nosso estudo. Na Seção 4, apresentamos os resultados do estudo. Fazemos algumas considerações finais na Seção 5.

2. Trabalhos relacionados

Ao longo das últimas décadas, várias abordagens foram utilizadas em IGNS. A maioria dos trabalhos utilizam a abordagem gerativa, principalmente no uso do algoritmo *Expectation–maximization – EM* [Baker 1979], que é utilizado para estimar a probabilidade de variáveis não observáveis (árvores sintáticas em IGNS). Nos últimos anos, a modelagem neural vem ganhando bastante espaço.

Ao longo das duas últimas décadas, o modelo DMV (*Dependency Model with Valence*) [Klein e Manning 2004] exerceu grande influência para gramática de dependência. A ideia por trás do modelo DMV está no controle de geração da árvore sintática, que, para cada ramo a ser gerado (relação de dependência), utiliza-se de distribuições de probabilidade para tomar decisões de quando gerar ($P_{STOP}(\neg STOP|h, dir, adj)$) e qual ramo gerar ($P_{CHOOSE}(a|h, dir)$). As variáveis *h*, *dir*, *a* e *dij* são respectivamente a cabeça da relação, a direção em que o argumento será gerado (direita ou esquerda), o argumento a ser gerado e se o argumento já foi gerado na árvore na direção *dir*. O DMV é um dos vários modelos que utilizam o EM. Este foi o primeiro trabalho a ultrapassar o *baseline* de ramificação direita (*right-branching*) [Headden III et al. 2009], sendo bastante utilizado, mesmo com quase duas décadas de existência [Yang et al. 2020].

Muitos trabalhos foram influenciados por [Klein e Manning 2004]. Um dos mais relevantes, [Headden III et al. 2009] estendeu o modelo DMV para aplicar uma abordagem Bayesiana, em vez de EM, utilizando uma gramática lexicalizada (cada nó da árvore sintática contém também informação sobre o léxico a que se refere). [Cohen e Smith 2009] optou por substituir a distribuição *Dirichlet* pela Logística, pois, apesar de a primeira ser mais fácil de treinar, ela não permite um meio explícito de forma flexível para calcular a covariância entre dois eventos, conforme descreve [Blei e Lafferty 2005]. O trabalho alcançou 42% de *Direct Dependency Accuracy – DDA* (quando considera a direção de geração da árvore sintática) no corpus WSJ∞, para sentenças de qualquer tamanho.

O trabalho de [Spitkovsky et al. 2010] obteve bons resultados a partir da aplicação de *curriculum learning* [Bengio et al. 2009], que inicia o treinamento com dados menos complexos e aumenta a complexidade dos dados até que toda a base de dados tenha sido utilizada. A mudança de complexidade contribui para que se reduzam as chances de cair em máximos locais (um dos problemas no uso de EM usado para problemas não convexos). Este trabalho obteve 45% de DDA no WSJ∞. Mais recentemente, [Han et al. 2019b] propôs o *Lexicalized Neural Dependency Model with Valence* (L-NDMV), um modelo lexicalizado que utiliza DMV com redes neurais. Esse trabalho constatou que, ao explorar características lexicais, a tarefa de IGNS ganha em desempenho. O L-NDMV foi o primeiro trabalho a ultrapassar a marca dos 60% de DDA no WSJ∞, enquanto que os trabalhos supervisionados ultrapassam a marca dos 95% em *Unlabeled Attachment Score – UAS* (quando não considera a direção de geração da árvore) [Lin et al. 2022].

[Yang et al. 2020] atingiu o estado da arte ao construir o modelo probabilístico com mais de um nível de distância de hierarquia (além de *filhos*, *pais* e *irmãos* também considera *avôs*, *netos* e *tios*, por exemplo) entre os nós da árvore. Outros trabalhos atingiram o estado da arte ao estender o modelo DMV com redes neurais [Han et al. 2019a] [Han et al. 2017] [Jiang et al. 2016]. Todos estes trabalhos utilizam algum tipo de informação léxica e redes neurais para contribuir com o desempenho. [Shen et al. 2021] usa o conceito de distância e altura sintática para segmentar a sentença em partes menores.

[Drozdov et al. 2019] aplica o algoritmo *Inside-Outside – IO*, que pode ser visto como uma instância do EM, em redes neurais.

Todos os trabalhos citados utilizam algum tipo de anotação no treinamento. Recentemente, [Pate e Johnson 2016] treinou o modelo DMV com milhões de palavras para induzir dependência sem uso de anotação. Uma vez que o modelo não utiliza categorias morfossintáticas, é utilizada inferência Bayesiana aplicada a gramáticas livres de contexto probabilísticas. Apesar de já existirem estudos anteriores que utilizavam apenas palavras como entrada para o modelo [Seginer 2007], estes eram apenas para constituintes.

O uso de IM, em específico, é algo que vem sendo relativamente pouco explorado, apesar de seu claro apelo para a tarefa. [Magerman e Marcus 1990] foi o primeiro trabalho a aplicar IM em IGS, mas foi recentemente que IM começou a ser aplicado em tarefas de IGNS. Em um trabalho recente, [Futrell et al. 2019] constatou que pares de palavras que têm relação sintática apresentam uma maior IM quando comparados a pares de palavras sem relação. Esta hipótese também foi aplicada em indução gramatical por [Hoover et al. 2021], que utilizou o modelo de língua pré-treinado para calcular informação mútua entre palavras considerando o contexto.

A seguir, detalhamos o método de [Futrell et al. 2019] e como o reproduzimos.

3. Método de indução gramatical

O trabalho de [Futrell et al. 2019] utilizou um corpus com 320 milhões de *tokens* anotados automaticamente. Os autores analisam três variáveis: palavras (*words*), categorias morfossintáticas (*pos*) e grupos lexicais (*lex*). A última variável é resultante de agrupamento. O trabalho propôs um agrupamento com os 60K *tokens* mais frequentes, incluindo *stopwords* e pontuação, a fim de ter uma dimensionalidade menor. Utilizando os vetores de 300 dimensões do modelo *Glove* para cada uma das 60K palavras, o trabalho agrupa os tokens em 300 grupos. Por exemplo, os *tokens* "carro", "carros", "automóvel" e "automóveis" fazem parte do mesmo grupo lexical. Para representar a variável *lex*, cada *token* no corpus é substituído pelo número do seu respectivo grupo.

A IM é calculada entre pares de palavras, categorias morfossintáticas e grupos lexicais. Para pares de palavras, por exemplo, na sentença "O menino chutou a bola", alguns dos pares possíveis são <o,chutou> e <chutou,bola>. O primeiro par não tem relação de dependência (indicada no experimento como *nondep*). O segundo par tem relação, conforme apresentado na Figura 1 (indicada como *dep*). [Futrell et al. 2019] quis também saber o desempenho de pares aleatórios. Ele descreveu estes pares como *permuted*. O mesmo é estabelecido para categorias morfossintáticas e grupos lexicais. Ao todo, para cada variável, 3 experimentos são realizados. O cálculo da IM ocorre entre o termo cabeça da relação *h* e o dependente *d*, cuja fórmula é apresentada abaixo. A fórmula calcula a probabilidade de haver uma relação entre duas variáveis, que podem ser palavras, categorias morfossintáticas, grupos lexicais e relações sintáticas (usadas neste trabalho).

$$IM = log \frac{P(h,d)}{P(h)P(d)}$$

Para avaliar seu modelo, [Futrell et al. 2019] utiliza dois *baselines*: pares permutados (*words perm*, *lex perm*, *pos perm*) e pares sem relação de dependência (*words nondep*,

lex nondep, *pos nondep*). [Futrell et al. 2019] utilizou estes dois baselines porque o primeiro considera uma relação aleatória na sentença, podendo ser de dependência (como em <o,menino>) ou não (<o,a>). Assim, o *baseline* permutado deve apresentar melhor desempenho do que o *baseline* sem dependência se existir maior informação mútua entre relações de dependência do que não dependência. Estes *baselines* são comparados com os resultados das 3 variáveis para relação de dependência (*words dep*, *lex dep*, *pos dep*).

Diferentemente de [Futrell et al. 2019], utilizamos dois corpora anotados por humanos, disponíveis na página do projeto UD: Bosque [Afonso et al. 2002] e Petrogold [de Souza et al. 2021]. Para o Bosque, as sentenças anotadas com Português do Brasil correspondem a 90K *tokens* e 4.205 sentenças originárias do CETENFolha [Linguateca 2023], construído com textos jornalísticos. O corpus PetroGold é formado por 19 teses e dissertações na área de óleo e gás, constituído por 232k *tokens* e 9k sentenças. O corpus Bosque contém 4,16% das sentenças com até 3 *tokens* e 8,3% das sentenças com mais de 40 *tokens*. O PetroGold contém 4,5% das sentenças com até 3 palavras e 24,7% com sentenças acima de 40 *tokens*. Além da distribuição diferente de tamanho de sentença, constatamos que os corpora apresentam também diferenças na distribuição de categorias morfossintáticas e funções sintáticas, apesar de essas diferenças serem pequenas.

Adotamos o método de [Futrell et al. 2019], mas utilizamos apenas anotações que foram produzidas por humanos, sem análise automática. Além disso, não selecionamos os *tokens* mais frequentes, uma vez que o tamanho do vocabulário dos corpora compilados é menor que o tamanho proposto por [Futrell et al. 2019]. Em vez disso, apenas utilizamos todo o vocabulário dos corpora que é representado no modelo Glove treinado para a língua portuguesa [Hartmann et al. 2017], totalizando um vocabulário de 21.428 palavras.

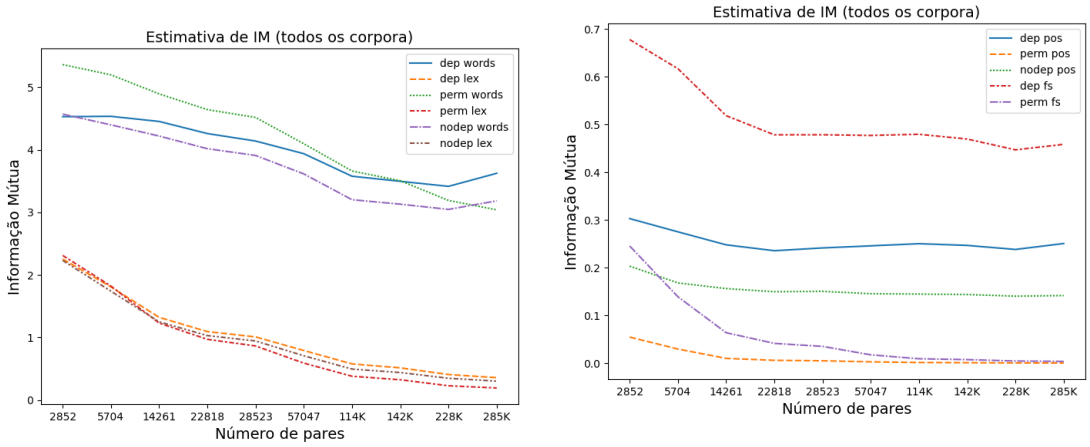
Neste trabalho, utilizamos *embeddings* de 50, 300 e 600 dimensões com base na análise realizada por [Hartmann et al. 2017] para gerar os grupos lexicais. Para definir os 300 grupos, [Futrell et al. 2019] utiliza uma matriz de similaridade. No entanto, [Futrell et al. 2019] não informa como essa matriz de similaridade foi construída. Deduzimos que a matriz foi construída usando similaridade de cosseno entre as *embeddings* de cada *token*. Devido às limitações computacionais, foram utilizadas apenas duas casas decimais para representar os valores na matriz de similaridade.

Além dos dois corpora isoladamente, também usamos a combinação deles. Para cada corpus, foram realizadas 3 execuções, uma para cada uma das 3 dimensões, contabilizando um total de 9 execuções. O código utilizado para a realização dos experimentos foi o mesmo disponibilizado por [Futrell et al. 2019].

4. Resultados

Na Figura 2, são apresentados os gráficos com os resultados para IM utilizando todos os corpora. Resolvemos utilizar os nomes originais das variáveis utilizados no trabalho do [Futrell et al. 2019] para facilitar a reprodução do estudo. Incluímos a variável *fs*, que representa as relações de dependência. Para o agrupamento, variável *lex*, foram utilizadas *embeddings* de 300 dimensões.

Na Figura 2(a), observa-se que, conforme o número de pares aumenta, decresce a informação mútua em todos os grupos, sem tendência de convergência para um valor específico. Este mesmo comportamento é observado no trabalho de [Futrell et al. 2019]

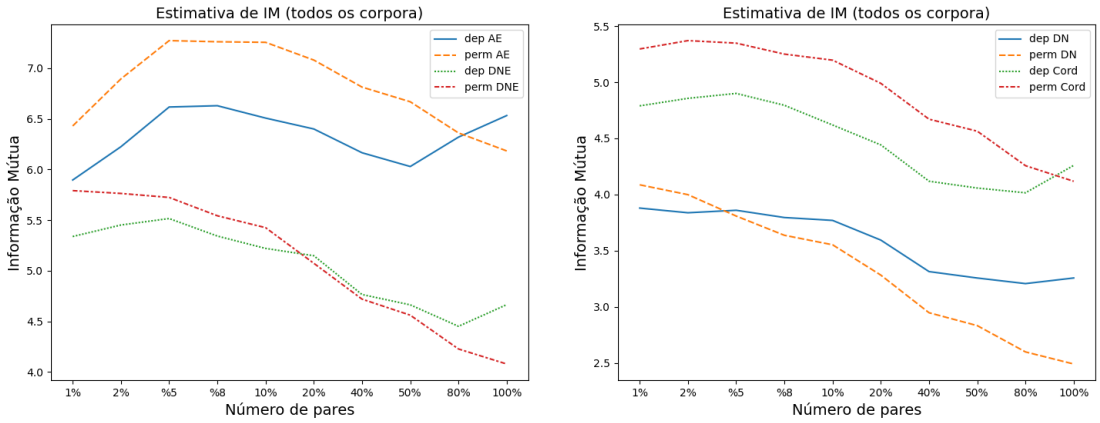


(a) IM de palavras e grupos léxico

(b) IM de grupos e categorias morfossintáticas (pos)

Figura 2. IM por número de pares usando 300 dimensões com todos os corpora

para a língua inglesa, mesmo utilizando um corpus dezenas de vezes maior que o nosso. Isso sugere que a IM entre pares que contenham relações de dependência pode seguir o mesmo padrão para diferentes línguas. Na Figura 2(b), para as variáveis *pos*, observa-se uma estabilidade, também apresentando comportamento similar ao trabalho de [Futrell et al. 2019]. Observa-se também uma IM maior para relações sintáticas do que para categorias morfossintáticas, mesmo com maior esparsidade no grupo de relações sintáticas.



(a) IM de relações sintáticas AE e DNE

(b) IM de relações sintáticas DN e Coord

Figura 3. IM por porcentagem dos pares utilizados com todos os corpora

Na Figura 3, são apresentados os resultados para IM de relações sintáticas separadas pelos grupos definidos pela UD: Argumentos Essenciais (AE) (*Core arguments*, que incluem, por exemplo, as relações sintáticas mais importantes da sentença, como sujeito e objetos), Dependentes Não Essenciais (DNE) (*Non-core dependents*, que incluem, por exemplo, relações de vocativo, modificadores adverbiais e verbos auxiliares), Dependentes Nominais (DN) (*Nominal dependents*, que incluem, por exemplo, modificadores de substantivos e de adjetivos) e coordenações (Coord) (*Coordination*, que incluem, por exemplo, fenômenos variados, como a coordenação por conjunções, expressões multipa-

lavra e relações especiais). Na Figura 3, podemos observar que, dos quatro grupos, AE apresenta a maior IM, assim como também é o único que não apresenta uma tendência de queda conforme o número de pares aumenta.

Durante os experimentos, observamos que o número de dimensões influencia na IM. Nos experimentos usando agrupamento, percebemos que, quanto maior o número de dimensões, menor será a IM, apesar de constatarmos um pequeno aumento de pouco mais de 5% da IM utilizando *embeddings* de 600 dimensões em comparação com *embeddings* de 300 dimensões. Não temos certeza do que pode ter causado esta variação, mas, uma vez que ocorreu uma redução de 50% de IM entre os grupos que utilizaram *embeddings* de 50 e 300, acreditamos que o resultado pode ser devido à alguma característica intrínseca aos *embeddings* utilizados. Não conseguimos identificar uma relação entre o tamanho da sentença e a IM. Um resumo dos resultados para cada corpus é apresentado na Tabela 1. Os dados sugerem que, quanto menor o número de pares, maior a informação mútua.

Tabela 1. Resumo das execuções para todos os corpora

Corpora	$\mu(\sigma)$ **palavras / sentença**	**Número de pares**	**IM dep words**	**IM nondep words**	**IM dep lex**	**IM nondep lex**
Bosque	21,5 (13,51)	70.938	4,638	4,016	0,352	0,301
PetroGold	30,0 (19,5)	221.987	3,275	2,893	0,213	0,187

Finalmente, realizamos experimentos usando as relações sintáticas (Tabela 2). Devido à diferença no número de pares entre as relações, realizamos os experimentos considerando 4 das relações mais relevantes: nsubj, obj, iobj e xcomp.

Tabela 2. Experimento utilizando informações sintáticas

Relações	**Número de pares**	**IM dep words**	**IM permuted**	**IM dep fs**	σ **fs**
nsubj	55.412	**7,488**	0,014	0,478	0,0057
obj	34.692	7,295	0,020	0,631	0.0061
iobj	677	5,026	**0,520**	**1,232**	**0,0159**
xcomp	8.357	5,897	0,088	0,940	0,0122

Os resultados apresentados na Tabela 2 demonstram que há uma diferença muito grande de IM entre relações de dependência e IM com relações permutadas. A relação *iobj* apresenta o melhor desempenho entre as demais relações quando se observa a direção da relação (IM de 1,232), provavelmente devido ao número pequeno de exemplos analisados, uma vez que a IM permutada é bastante alta e o desvio padrão também. No caso da relação *nsubj*, a terceira coluna representa a IM das palavras que fazem parte dessa relação de dependência (sem considerar a direção da relação). Percebe-se uma alta IM nesta categoria, provavelmente devido às características sintáticas do *nsubj*. Como ilustração de pares com alta IM, as maiores IM encontradas para *obj* foram para os pares **computadores – comprei** (com valor 0,03197), **anos – há** (0,01870) e **-se – trata** (0,01336).

5. Considerações finais

Reproduzimos o trabalho de [Futrell et al. 2019] usando corpora da língua portuguesa. Apesar de o tamanho dos corpora usados por [Futrell et al. 2019] e os usados neste trabalho serem bem diferentes, constatamos tendência de comportamento similares, com algumas pequenas diferenças, sugerindo que existe um padrão de comportamento mesmo em línguas pertencentes à famílias linguísticas diferentes. Diferentemente do estudo publicado por [Futrell et al. 2019], que anotou milhões de palavras automaticamente, utilizamos apenas anotações sintáticas de referência produzidas por humanos. Essas variações tornam inconclusivas comparações diretas entre os trabalhos. Além disso, [Futrell et al. 2019] não informa como foi construída a matriz de similaridades.

Trabalhos futuros incluem aplicar este experimento a outros corpora anotados, como o Porttinari [Pardo et al. 2021], e realizar um estudo mais aprofundado sobre a influência nos resultados dos corpora, assim como das *embeddings* utilizadas.

Agradecimentos
Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Afonso, S., Bick, E., Haber, R., e Santos, D. (2002). Floresta sinta(c)tica: A treebank for portuguese. In the *Proceedings of the Conference on Language Resources and Evaluation* (LREC), 1698–1703.

Baker, J. K. (1979). Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 132–132.

Bannard, C., Lieven, E., e Tomasello, M. (2009). Modeling children's early grammatical knowledge. In the *Proceedings of the National Academy of Sciences* (PNAS), 17284–17289.

Bengio, Y., Louradour, J., Collobert, R., e Weston, J. (2009). Curriculum learning. In the *Proceedings of the 26th Annual International Conference on Machine Learning* (ICML), 41–48.

Blei, D. M. e Lafferty, J. D. (2005). Correlated topic models. In the *Proceedings of Advances in Neural Information Processing Systems* (NIPS), 147–154.

Bod, R. (2007). Is the end of supervised parsing in sight? In the *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (ACL), 400–407.

Bresnan, J., Asudeh, A., Toivonen, I., e Wechsler, S. (2015). Lexical-functional syntax. John Wiley & Sons.

Chomsky, N. (2014). Aspects of the Theory of Syntax, volume 11. MIT press.

Cohen, S. B. e Smith, N. A. (2009). Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics* (NAACL), 74–82.

da Costa, P. B. e Kepler, F. N. (2014). Semi-supervised parsing of portuguese. In the *Proceedings of the Computational Processing of the Portuguese Language - 11th International Conference* (PROPOR), 102–107.
Dahl, V., Bel-Enguix, G., Tirado, V., e Miralles, J. E. (2023). Grammar induction for under-resourced languages: The case of ch'ol. In the *Proceedings of the Analysis, Verification and Transformation for Declarative Programming and Intelligent Systems - Essays Dedicated to Manuel Hermenegildo on the Occasion of His 60th Birthday*, 113–132.
de Marneffe, M.-C., Manning, C. D., Nivre, J., e Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 255–308.
de Souza, E., Silveira, A., Cavalcanti, T., Castro, M. C., e Freitas, C. (2021). Petrogold corpus padrão ouro para o domínio do petroleo. *In Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (STIL), 29–38.
Drozdov, A., Verga, P., Yadav, M., Iyyer, M., e McCallum, A. (2019). Unsupervised latent tree induction with deep inside-outside recursive autoencoders. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics* (NAACL), 1129-1141.
Klein, D. e Manning, C. D. (2002). A generative constituent-context model for improved grammar induction. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL), 128–135.
Futrell, R., Qian, P., Gibson, E., Fedorenko, E., e Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In the *Proceedings of the fifth international conference on dependency linguistics* (depling), 3–13.
Han, W., Jiang, Y., e Tu, K. (2017). Dependency grammar induction with neural lexicalization and big training data. In the *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 1683–1688.
Han, W., Jiang, Y., e Tu, K. (2019a). Enhancing unsupervised generative dependency parser with contextual information. In the *Proceedings of the 57th Conference of the Association for Computational Linguistics* (ACL), 5315–5325.
Han, W., Jiang, Y., e Tu, K. (2019b). Lexicalized neural unsupervised dependency parsing. *Neurocomputing*, 105–115.
Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., e Alu´ısio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In the *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology* (STIL), 122–131.
Hoover, J. L., Du, W., Sordoni, A., e O'Donnell, T. J. (2021). Linguistic dependencies and statistical dependence. In the *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2941–2963.
Headden III, W. P., Johnson, M., e McClosky, D. (2009). Improving unsupervised dependency parsing with richer contexts and smoothing. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics* (NAACL), 101–109.
Jiang, Y., Han, W., e Tu, K. (2016). Unsupervised neural dependency parsing. In the *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 763–771.
Klein, D. e Manning, C. D. (2004). Corpus-based induction of syntactic structure:

Models of dependency and constituency. In the *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (ACL), 478–485.
Lin, B., Yao, Z., Shi, J., Cao, S., Tang, B., Li, S., Luo, Y., Li, J., e Hou, L. (2022). Dependency parsing via sequence generation. *Findings of the Association for Computational Linguistics*, 7339–7353.
Linguateca (2023). Cetem publico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Linguateca,http://www.linguateca.pt/CETENFolha/, última visita: Junho de 2023.
Magerman, D. M. e Marcus, M.a P. (1990). Parsing a natural language using mutual information statistics. In the *Proceedings of the 8th National Conference on Artificial Intelligence* (AAAI), 984–989.
Pardo, T. A. S., Duran, M. S., Lopes, L., Felippo, A. d., Roman, N. T., e Nunes, M. d. G. V. (2021). Porttinari: a large multi-genre treebank for brazilian portuguese. In the *Proceedings of the XIII Symposium in Information and Human Language* (STIL), 1–10.
Pate, J. K. e Johnson, M. (2016). Grammar induction from (lots of) words alone. In the *Proceedings of 26th International Conference on Computational Linguistics* (COLING), 23–32.
Seginer, Y. (2007). Fast unsupervised incremental parsing. In the *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (ACL), 384–391
Shen, Y., Tay, Y., Zheng, C., Bahri, D., Metzler, D., e Courville, A. C. (2021). Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In the *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (ACL/IJNLP), 7196–7209.
Spitkovsky, V. I., Alshawi, H., e Jurafsky, D. (2010). From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In the *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics* (NAACL), 751–759.
Spitkovsky, V. I., Alshawi, H., e Jurafsky, D. (2013). Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In the *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 1983–1995.
Stevenson, A. e Cordy, J. R. (2014). A survey of grammatical inference in software engineering. *Science of Computer Programming*, 444–459.
Theodor, C. C. e Siebert-Cole, E. (2020). Family tree of languages. https://www.researchgate.net/publication/342850691 TREES of LANGUAGES 2022, ultima visita:junho 2023.
Unold, O., Gabor, M., e Dyrka, W. (2020). Unsupervised grammar induction for revealing the internal structure of protein sequence motifs. In the *Proceedings of Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine* (AIME), 299–309.
Yang, S., Jiang, Y., Han, W., e Tu, K. (2020). Second-order unsupervised neural dependency parsing. In the *Proceedings of the 28th International Conference on Computational Linguistics* (COLING), 3911–3924