Análise das Interações Sociais em Comunidades Online de Aprendizado de Idiomas: um estudo de caso no Reddit*

Rafael Sales Medina, Ana Paula Couto da Silva, Fabricio Murai

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{rafael.medina, ana.coutosilva, murai}@dcc.ufmg.br

Abstract. *Reddit is a online social network in which users can share information about mutual interests in specific communities (subreddits). Recently, communities focused on language learning have gained much popularity among users. These subreddits enable users to interact, regardless of their proficiency level on a specific language. Typical interactions include answering questions and sharing tips for facilitating the learning process. In this paper, we analyze four of these communities: EnglishLearning, French, German and Spanish. This analysis focuses on interactions between users, how discussion revolves around threads and linguistic traits of users belonging to different proficiency levels. Moreover, we highlight similarities and differences among these communities.*

Resumo. *Reddit é uma rede social online na qual os usuários podem trocar informação sobre interesses comuns em comunidades específicas (subreddits). Recentemente, comunidades voltadas para o aprendizado de idiomas vêm ganhando popularidade. Esses subreddits permitem que usuários interajam, independentemente do seu nível de proficiência. Interações típicas incluem a resolução de dúvidas e a troca de dicas para facilitar o aprendizado. Neste trabalho, analisamos quatro comunidades: EnglishLearning, French, German e Spanish. Esta análise foca em interações entre usuários, como as discussões se desenrolam em torno das threads e os traços linguísticos dos usuários que pertencem a diferentes níveis de proficiência. Além disso, ressaltamos as semelhanças e diferenças entre essas comunidades.*

1. Introdução

Nos últimos anos, as redes sociais online têm sido utilizadas para diversas finalidades, como manter contato com amigos antigos, fazer novas amizades, compartilhar atualizações [Joinson 2008] e até mesmo para formação de grupos de apoio virtuais, como aqueles voltados para emagrecimento [Pappa et al. 2017, Cunha et al. 2016] e para vítimas de abuso sexual [Andalibi et al. 2016].

Uma rede que permite o contato entre pessoas em torno de um tema de interesse comum é o Reddit[1], um site de fóruns com características de redes sociais onde os membros compartilham suas experiências e dúvidas sobre os mais diversos assuntos. O Reddit

*Este trabalho foi parcialmente apoiado por recursos do CNPq e do projeto FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, número de processo APQ-01400-14.

[1]http://www.reddit.com

é organizado em comunidades (subreddits) e, em 2017, era formado por 1.204.126 comunidades com 900 milhões de comentários. Existem subreddits específicos para discutir programas de televisão[2], jogos de computador[3] e até mesmo para compartilhamento de tópicos relacionadas à saúde, como dicas de emagrecimento[4] e suporte mútuo em relação a problemas de saúde mental, como aqueles estudados em [De Choudhury and De 2014].

Em particular, comunidades específicas voltadas para o aprendizado de idiomas estrangeiros vem ganhando muita popularidade, como EnglishLearning[5] (será abreviada como English), French[6], German[7] e Spanish[8]. Esses subreddits permitem que pessoas com interesse em certo idioma interajam, compartilhando dúvidas, sugestões e dicas, tornando o processo de aprendizado mais dinâmico e interessante.

Como o aprendizado online e as redes sociais online atraem cada vez mais a atenção das pessoas no mundo inteiro, neste trabalho caracterizamos as atividades e interações dos usuários de subreddits voltados para o aprendizado de idiomas. Mais precisamente, nossas análises têm como objetivo responder as seguintes perguntas de pesquisa:

- **QP1:** As interações entre usuários em um subreddit são semelhantes àquelas em uma rede social tradicional?
- **QP2:** Como as publicações em um subreddit estão distribuídas em relação às threads?
- **QP3:** Existem diferenças linguísticas no texto de usuários com diferentes níveis de proficiência?

Para responder estas questões, modelamos a rede de usuários e suas interações dentro de um subreddit como um grafo, seguindo trabalhos recentes da literatura [Pappa et al. 2017]. Métricas como o *closeness*, o grau médio de entrada e saída e o coeficiente de clusterização são usadas para analisar os padrões de comportamento dos usuários que participam dos subreddits analisados. As publicações feitas pelos usuários são analisadas através da definição de árvores de discussão, que permitem investigar características importantes como a profundidade e a largura das *threads*. Adicionalmente, utilizamos a ferramenta LIWC (*Linguistic Inquiry and Word Count*)[9] para identificar diferenças linguísticas nos *posts* de membros do subreddit *German* associados a diferentes níveis de proficiência. Nossos resultados são importantes para a definição de novos modelos de aprendizado de idiomas apoiado por tecnologia considerando, por exemplo, a evolução da proficiência e o perfil de interação de usuários participantes destes subreddits.

Este artigo está organizado da seguinte forma: a Seção 2 descreve os principais trabalhos relacionados; a Seção 3 detalha os dados e métodos utilizados; os resultados da análise dos subreddits são apresentados na Seção 4; as implicações deste trabalho e trabalhos futuros são discutidos na Seção 5.

[2]http://www.reddit.com/r/rupaulsdragrace
[3]http://www.reddit.com/r/thesims
[4]http://www.reddit.com/r/loseit
[5]http://www.reddit.com/r/EnglishLearning/
[6]http://www.reddit.com/r/French/
[7]http://www.reddit.com/r/German/
[8]http://www.reddit.com/r/Spanish/
[9]http://liwc.wpengine.com/

2. Trabalhos Relacionados

Há pelo menos duas décadas é possível encontrar trabalhos voltados para o aprendizado de idiomas apoiado por tecnologias [Zhao 1996, Warschauer and Healey 1998]. Esse campo de estudos é chamado de *Computer Assisted Language Learning* (CALL) [Levy 1997]. CALL engloba quaisquer tipos de aplicações que possam auxiliar no aprendizado de um idioma estrangeiro, como as redes sociais online, que são o foco deste trabalho. O interesse dos pesquisadores na área de redes sociais é relativamente recente, como descrito por [Zourou 2012]. Neste mesmo trabalho, a autora discute o estado da arte em relação ao uso de mídias sociais para o ensino de idiomas, mas sem especificar uma rede ou linguagem. Ela demonstra que as redes têm influência positiva no ensino e são bastante utilizadas por fomentarem participação dos usuários.

Os autores de [Arnold and Paulus 2010] analisam, sob a visão de estudantes, de um instrutor e de um observador externo, um caso prático em que uma turma real de aprendizado de idioma utilizou o sistema Ning, voltado para a criação de comunidades sociais. Este trabalho conclui, sob a visão do instrutor de idioma, que a utilização da comunidade social para discussão teve resultado positivo no ensino.

Sob a ótica de pessoas que estão aprendendo um novo idioma, a pesquisa realizada em [Lin et al. 2016] analisa o comportamento e desenvolvimento de usuários de redes específicas para aprendizado de idiomas. O estudo, feito por meio de questionários e acompanhamento de estudos de casos, conclui que as redes sociais voltadas para o aprendizado de idiomas apresentam resultados positivos, mas também limitações. Além disso, conclui-se que para que os usuários alcancem o sucesso esperado, é necessário que as redes ofereçam apoio, orientação e atividades bem estruturadas, de maneira a promover o engajamento e interação dos usuários.

Apesar de existirem abordagens à utilização de redes sociais para o auxílio do aprendizado de idiomas, os trabalhos citados anteriormente focam em redes criadas especificamente para o escopo de ensino e aprendizado. O nosso trabalho analisa como os usuários do Reddit, que é uma rede social composta por comunidades criadas em torno dos mais diversos tópicos, pode auxiliar no aprendizado de um determinado idioma. A partir do estudo as interações dos usuários, dos seus interesses e como os mesmos se organizam em torno de tópicos em comum, fóruns e comunidades poderão ser criados, visando um resultado mais positivo no aprendizado de novos de idiomas. Segundo o nosso conhecimento, este é o primeiro trabalho a investigar comunidades no Reddit que focam no aprendizado de idiomas.

3. Metodologia

Apresentamos abaixo os métodos utilizados neste trabalho para o estudo de subreddits focados no aprendizado de idiomas. Em particular, descrevemos as técnicas usadas para coleta e extração dos dados, modelagem da interação entre usuários e análise textual.

3.1. Coleta e extração dos *datasets*

Os dados do Reddit analisados neste trabalho foram obtidos a partir de *dumps* disponíveis na Web[10]. Coletamos todas as atividades realizadas por usuários nos subreddits English,

[10]http://files.pushshift.io/reddit/

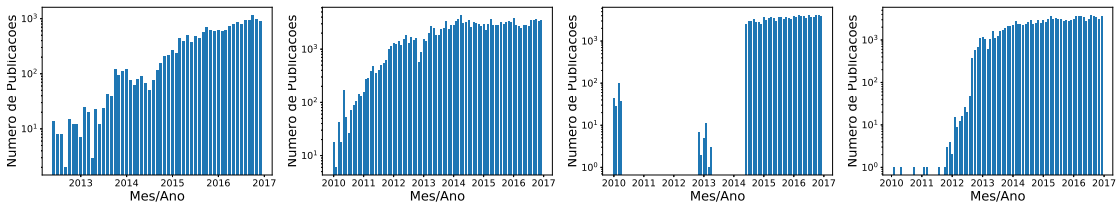
French, German e Spanish entre janeiro de 2010 e dezembro de 2016. As atividades consistem em *posts* e comentários feitos pelos usuários das comunidades.

A Tabela 1 apresenta o número de usuários, o total de *posts* e comentários observados durante este intervalo, bem como os valores médios por usuário, para cada um dos subreddits. A comunidade voltada para ensino de inglês é aquela que tem o menor número de usuários. Um dos motivos que provavelmente contribui para que isto aconteça é que a maior parte dos visitantes do Reddit é de países de língua inglesa. Em dezembro de 2017, os três países seguintes correspondiam a mais de metade dos usuários ativos[11]: Estados Unidos (39,79%), Inglaterra (7,16%) e Austrália (3,46%).

	EnglishLearning	French	German	Spanish
Usuários	3.737	18.113	12.235	13.648
Posts	5.493	17.329	12.441	14.295
Comentários	11.967	145.700	96.265	113.843
Média de *posts* por usuário	1,47	0,96	1,02	1,05
Média de comentários por usuário	3,20	8,04	7,87	8,34

Tabela 1. Estatísticas dos subreddits no período entre 2010 e 2016.

A Figura 1 mostra o volume total de publicações nos subreddits a cada mês. Observa-se que os subreddits English e French já ganhavam popularidade desde 2010. No German houve algumas rajadas de atividade em 2010 e 2013, tendo-se observado um salto abrupto no volume de publicações durante 2014. No Spanish também observou-se um salto, embora menos pronunciado, em 2012. Em todos os casos, o volume de atividades em 2016 se manteve relativamente constante.



(a) English (b) French (c) German (d) Spanish

Figura 1. Volume mensal de publicações (*posts* e comentarios).

3.2. Modelo de Interação entre Usuários

Para cada subreddit, modelamos as interações entre os usuários como um grafo direcionado ponderado $G_d = (V, E_d, W_d)$, onde V é um conjunto de vértices, E_d é um conjunto de arestas e W_d é uma função que mapeia cada aresta $e \in E_d$ a um peso $W_d(e) \in \mathbb{R}$. Cada vértice representa um usuário que tenha publicado um *post* ou comentário no subreddit, e cada aresta indica uma interação entre dois usuários. As arestas são direcionadas: v aponta para u se o vértice v respondeu a um *post* ou comentário de um vértice u. Para cada aresta $e = (i, j) \in E_d$, o peso $W_d(e)$ é igual ao número de interações de i com j.

[11] https://www.statista.com/statistics/325144/reddit-global-active-user-distribution

Definimos também o grafo não-direcionado ponderado $G = (V, E, W)$ induzido por G_d ao tornarmos as arestas em E_d não-direcionadas e fazermos $W(e) = W_d(e) + W_d(e')$, para $e = (i, j)$ e $e' = (j, i)$.

Utilizamos o grafo direcionado G_d para caracterizar um subreddit quanto à distribuição do volume de atividades dos seus usuários medido em termos (i) dos graus de entrada e (ii) de saída e (iii) do *closeness*. O *closeness* é uma métrica clássica de centralidade em redes que tenta capturar a importância relativa dos nós a partir da estrutura do grafo [Newman 2011]. Além disso, usamos também o grafo não-direcionado G para calcular a distribuição do coeficiente de clusterização dos nós.

A Tabela 3.2 apresenta os dados básicos dos grafos de interação G_d, incluindo a quantidade de vértices (usuários) e arestas, de componentes conexos e o tamanho do maior componente conexo de cada rede.

	EnglishLearning	French	German	Spanish
Vértices	3.737	18.113	12.235	13.648
Arestas	8.881	95.152	66.309	74.108
Número de componentes conexos	888	1.435	776	12.14
Vértices no maior componente	2.804	16.607	11.425	12.398
Arestas no maior componente	6.516	72.231	49.814	56.463

Tabela 2. Características dos grafos de interação.

3.3. Análise dos *Posts* e Comentários

Cada *thread* em um subreddit pode ser vista como uma árvore iniciada por um post (nó raiz) que pode ser respondido diretamente por comentários. Cada comentário pode, por sua vez, ser respondido por outros comentários. Denominamos **árvores de discussão** as árvores que reconstruímos ao mapearmos cada comentário em um subreddit ao seu "nó pai". Iremos calcular a profundidade e a largura destas árvores a fim de identificar os tópicos que despertam o maior interesse dos usuários. Além disso, considerando os *timestamps* associados aos nós, iremos mensurar a taxa de crescimento destas árvores.

Adicionalmente é possível analisar o texto das publicações com o auxílio da ferramenta LIWC (*Linguistic Inquiry and Word Count*)[12], que realiza a análise automatizada de textos em diversas línguas e os classifica em diferentes categorias. Essa análise permite relacionar características linguísticas do texto, baseadas nos resultados do LIWC, com a proficiência indicada pelo usuário em seu perfil.

4. Análise dos resultados

Nesta seção descrevemos os resultados obtidos através dos métodos descritos na Seção 3 e explicar como eles respondem as questões de pesquisa levantadas neste trabalho.

4.1. QP1: As interações entre usuários nesses subreddits são semelhantes àquelas em uma rede social tradicional?

A partir de G_d, calculamos a distribuição conjunta de graus de entrada e de saída dos vértices. O grau de entrada de um vértice é a quantidade de comentários que o usuário

[12]http://liwc.wpengine.com/

correspondente recebeu em suas publicações. Por outro lado, o grau de saída de um vértice é a quantidade de publicações feitas por um usuário. A distribuição conjunta é mostrada na Figura 2 através de um mapa de calor em escala log-log, onde a cor do ponto (i, j) indica a fração de vértices em G_d com grau de entrada $i - 1$ e grau de saída $j - 1$.

Para todos os subreddits, observamos uma distribuição de cauda pesada em relação a ambas as distribuições marginais, além de uma forte correlação entre grau de entrada e grau de saída. Em sua grande maioria, usuários fazem e recebem poucos comentários, enquanto alguns poucos que fazem muitos comentários também recebem muitas respostas em suas publicações. Em particular, o subreddit English se destaca por apresentar pouquíssimos nós com graus de entrada ou saída maiores que 10^2, o que pode ser explicado por ter menos usuários que outras comunidades.

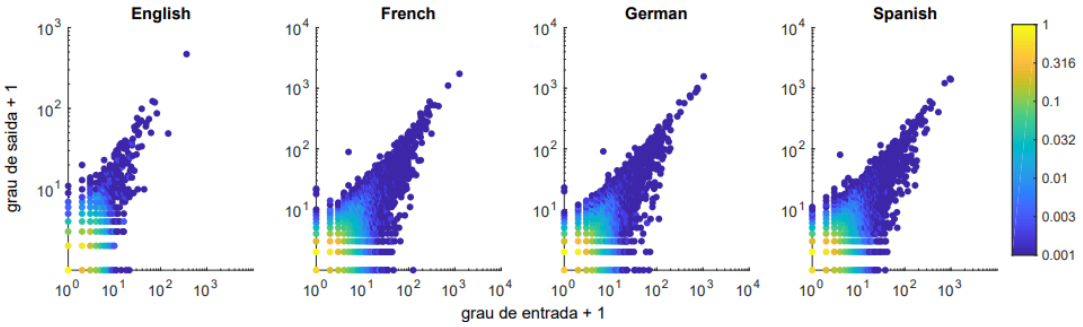


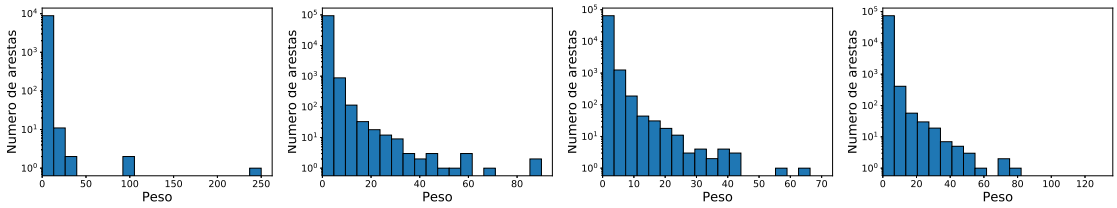
Figura 2. Distribuição conjunta do grau de entrada e saída. Cor indica a fração de participantes com dado grau de entrada e saída.

Embora a distribuição de graus de entrada e saída em um subreddit tenham cauda pesada assim como grande parte das redes sociais online, é natural ponderar se os grafos G_d e G exibem características locais semelhantes a essas redes. Para responder esta questão, mensuramos algumas destas características através das seguintes métricas:

- Pesos das arestas: medem a intensidade da interação entre pares de usuários. Definido como número de interações entre um par durante o intervalo considerado.
- Centralidade de *closeness*: mede o quão próximo um vértice está dos outros. Definido como o inverso da soma das distâncias entre um vértice e cada um dos outros vértices alcançáveis.
- Coeficiente de clusterização: mede o quão conectados estão os vizinhos de um nó. Definido como a fração de arestas existentes entre vizinhos de um nó dentre o máximo possível.

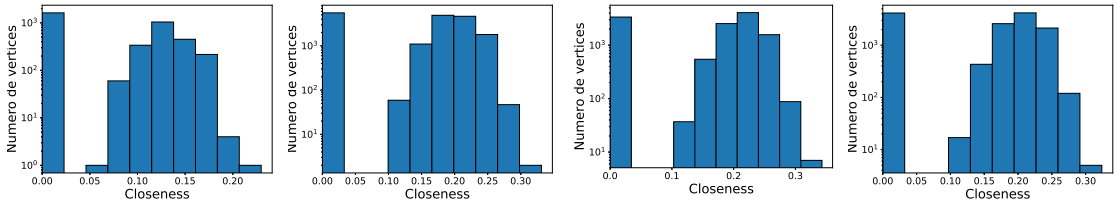
A Figura 3 mostra os histogramas de pesos nas arestas obtidos para cada subreddit. Observamos que a maioria arestas tem peso 1, indicando apenas uma interação entre dois usuários. Poucas arestas possuem peso elevado (p. ex., acima de 20), o que indica altos níveis de interação entre poucos pares de vértices.

A Figura 4 mostra os histogramas da distribuição do *closeness* para cada subreddit. Observam-se valores baixos, indicando distâncias longas entre vértices. Isto corrobora a intuição de que os usuários não têm interesse específico em conhecer pessoas, e sim em compartilhar conteúdo de interesse à rede. As conexões são formadas conforme os tópicos de interesse comum são compartilhados.



(a) English (b) French (c) German (d) Spanish

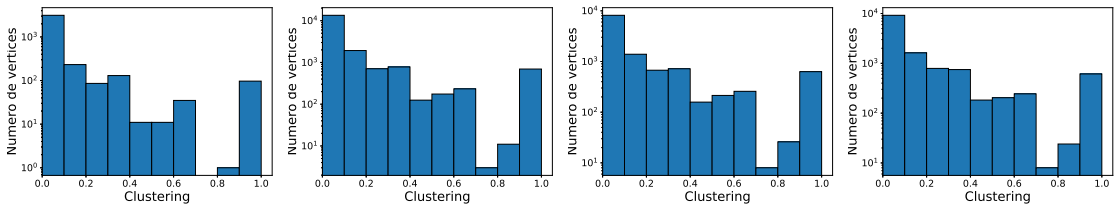
Figura 3. Distribuição do peso das arestas.



(a) English (b) French (c) German (d) Spanish

Figura 4. Distribuição do *closeness*.

A Figura 5 apresenta os histogramas do coeficiente de clusterização nas redes. É possível observar que um número elevado de usuários apresenta este valor igual a zero e muitos apresentam um valor baixo para esta métrica, o que indica que as redes são esparsas e não apresentam muita formação de triângulos, como no caso de redes sociais tradicionais. Isso reflete novamente na principal característica do Reddit, que é centrado em conteúdo em vez de amizade entre usuários.



(a) English (b) French (c) German (d) Spanish

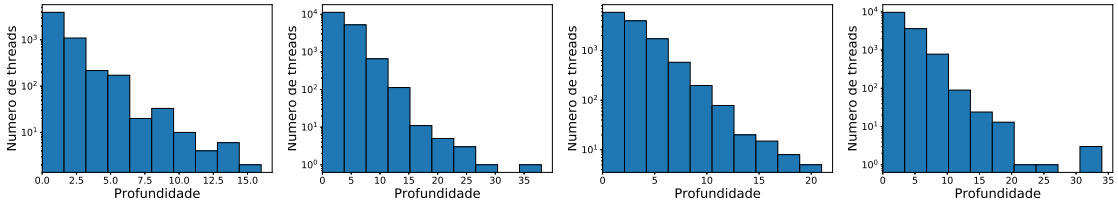
Figura 5. Distribuição do coeficiente de clusterização.

4.2. QP2: Como as publicações em um subreddit estão distribuídas em relação às *threads*?

Além da análise dos grafos, também foram analisados os *posts* e comentários dos subreddits através das árvores de discussão. Primeiramente, investigamos a distribuição da profundidade destas árvores. A profundidade é um indicador da progressão de discussões, já que uma árvore com muitos níveis indica que na *thread* correspondente houve pelo menos uma longa cadeia de comentários. A Figura 6 mostra a distribuição das profundidades para cada um dos subreddits.

Observa-se que a maioria das postagens tem a profundidade baixa e que são poucas as postagens que terminam em discussões longas. As árvores de profundidade 1 são aquelas cujos *posts* não receberam nenhum comentário. Por conveniência, mostramos a

fração de *posts* respondidos em cada subreddit na Tabela 3. Os subreddits French, German e Spanish apresentam altos índices de respostas às dúvidas compartilhadas pelos membros.



(a) English (b) French (c) German (d) Spanish

Figura 6. Distribuição da profundidade das árvores de discussão.

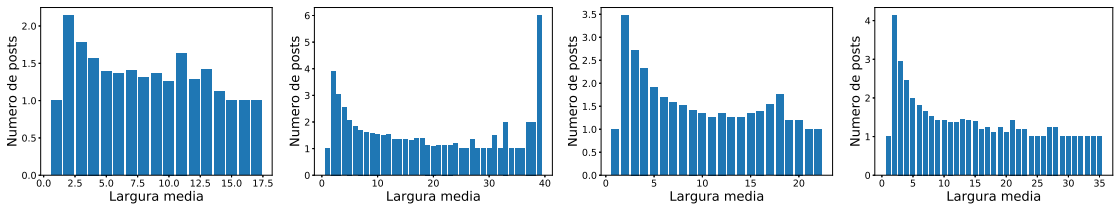
	English	French	German	Spanish
Com resposta	53,55%	83,88%	88,15%	81,56%
Sem resposta	46,45%	16,12%	11,85%	18,44%

Tabela 3. Porcentagem de *posts* com e sem respostas.

Para tentarmos compreender melhor o que leva uma *thread* a se prolongar por muitos níveis, foi realizada uma análise qualitativa daquelas com a maior profundidade considerando cada subreddit. Observamos que no subreddit *German*, a maior árvore é voltada ao compartilhamento de dicas de pronúncia, enquanto no *French* o tema está relacionado à dicas para melhora do vocabulário. Para o *English*, a *thread* de maior profundidade discute as diferenças entre o inglês coloquial e forma culta.

Na comunidade *Spanish* a árvore de discussão mais longa tem um tema menos voltado para proficiência e mais para a vivência, dado que a discussão gira em torno de como a imersão na cultura estrangeira é uma das melhores maneiras de se aprender um novo idioma. Esta análise sugere indícios de que *threads* de maior engajamento dos usuários tendem a ter como assunto principal conselhos para melhora da proficiência de um aluno.

Em seguida, investigamos a largura média das árvores da discussão, ou seja, a quantidade média de comentários que cada *thread* teve em cada nível de profundidade. A Figura 7 mostra a distribuição média da largura por nível de profundidade nos subreddits.



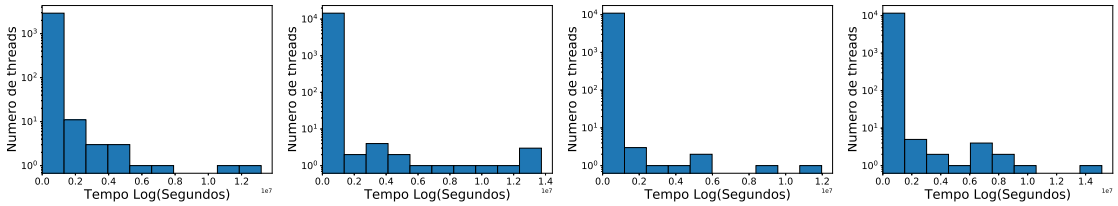
(a) English (b) French (c) German (d) Spanish

Figura 7. Largura média por nível das árvores de discussão (condicionado em profundidade $>$ nível).

Para permitir uma melhor comparação, na Figura 7 o primeiro nível indica a quantidade de *posts* nos subreddits. É possível observar que o engajamento dos usuários é

muito elevado nos primeiros níveis de profundidade, ou seja, os usuários respondem diretamente ao post ou aos primeiros comentários. Isso pode indicar que as discussões não se prolongam por muitos comentários e são poucas as *threads* em que a interação dos usuários ocorre por muito tempo.

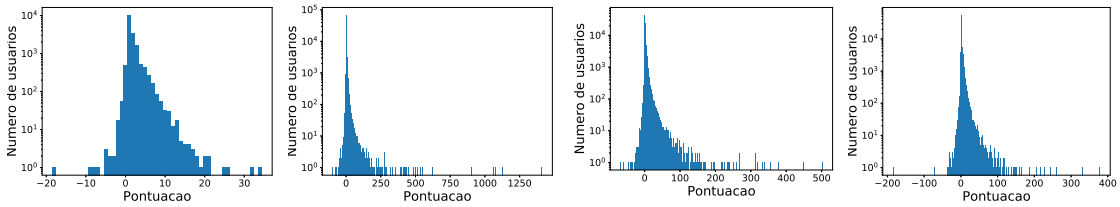
Outra métrica importante de engajamento dos usuários é o tempo decorrido até que um post seja respondido pela primeira vez. A Figura 8 mostra que, dentre os *posts* que receberam uma resposta, a grande maioria foi respondida rapidamente e que poucos ficaram dias até serem respondidos.



(a) English (b) French (c) German (d) Spanish

Figura 8. Distribuição do tempo decorrido até a primeira resposta em uma *thread*.

Uma outra métrica que pode ser avaliada está relacionada à pontuação dos *posts*, calculada pela diferença entre o número de *upvotes* e *downvotes*. O próprio Reddit ordena as publicações pela pontuação final: quanto maior, mais destaque tem o *post* e, consequentemente, mais usuários poderão interagir nesta *thread*. A Figura 9 mostra a distribuição da pontuação dos *posts* por subreddit.



(a) English (b) French (c) German (d) Spanish

Figura 9. Distribuição da pontuação das *threads*.

É possível observar que a pontuação possui valores baixos, em torno de zero. Isso quer dizer que os usuários não têm muito costume de votar nos *posts*, o que reforça a ideia de que as interações são voltadas para o conteúdo e não aprofundam nas discussões.

4.3. QP3: Existem diferenças linguísticas no texto de usuários com diferentes níveis de proficiência?

Para esta questão, iremos utilizar apenas o subreddit German, pois este solicita explicitamente aos usuários que adicionem *tags* (*flairs*) de proficiência a seus perfis. Consequentemente, o German possui um número muito grande de usuários que indicam seu nível de fluência quando comparado às outras comunidades.

O número de usuários do subreddit German com as *tags* de proficiência Iniciante, Intermediário, Avançado e Nativo é, respectivamente, 774, 583, 197 e 896. Um total de 9785 perfis de usuários não possui nenhum desses *tags*. A partir dos *posts* associados a

usuários com certo nível de proficiência, fizemos uma análise textual usando a ferramenta LIWC. Avaliamos quatro propriedades: (i) o número de palavras no texto; (ii), o número de palavras por frase; (iii) a quantidade de palavras com mais de 6 letras; e (iv) a quantidade de palavras reconhecidas pelo dicionário do LIWC. Essas duas últimas propriedades assumem valores de 0 a 100.

A Tabela 4 mostra as médias e medianas de cada propriedade para cada grupo de *posts*. Os valores mais elevados para cada propriedade aparecem em negrito. Observa-se que usuários com alemão avançado ou de língua nativa tendem a usar mais palavras nos *posts*. Além disso, estes usuários tendem a usar mais palavras por frase. Considerando "Nativo" como mais proficiente que "Avançado", pode-se observar que o tamanho das palavras utilizadas cresce com a proficiência e que o uso de palavras do dicionário diminui com ela (possivelmente dando lugar a gírias e expressões idiomáticas).

		Iniciante	**Intermediário**	**Avançado**	**Nativo**
Contagem de palavras	Mediana	17,00	18,00	**29,00**	25,00
	Média	44,36	40.44	54,12	**58,77**
Palavras por frase	Mediana	7,00	8,30	**11,00**	10,25
	Média	7,67	8,98	**11,63**	11,44
Palavras grandes	Mediana	16,67	19,35	21,01	**21,05**
	Média	17,29	19,61	21,25	**22,39**
Palavras do dicionário	Mediana	**64,71**	64,58	62,96	59,04
	Média	**66,32**	65,46	63,27	60,67

Tabela 4. Resultados da análise textual utilizando a ferramenta LIWC.

Os resultados da análise textual das publicações indicam que existem diferenças na forma como as pessoas em diferentes níveis de proficiência escrevem no subreddit. Esse fato pode ser utilizado para estimar a proficiência do usuário, quando esta é desconhecida. Contudo, esta análise possui duas limitações: a proficiência é auto-declarada e, portanto, as médias e medianas podem estar super- ou subestimadas; apesar da comunidade solicitar o uso das *tags* explicitamente, 80% dos usuários não as tem em seus perfis, podendo o viés daqueles que as tem ser significativo sobre as métricas estudadas.

5. Conclusão

Neste trabalho analisamos as interações sociais em quatro comunidades do Reddit voltadas para o aprendizado de idiomas: English, French, German e Spanish. Para isto, coletamos *posts*, comentários, *upvotes*, *downvotes* realizados por usuários nestes subreddits entre janeiro de 2010 e dezembro de 2016. A partir destes dados, geramos um grafo de interação entre usuários e árvores de discussão, utilizados para responder três perguntas de pesquisa.

Em relação à **QP1**, concluímos que as interações dentro dessas comunidades diferem daquelas em redes sociais tradicionais. Embora a distribuição de graus de entrada e saída tenham cauda pesada, o volume de interação entre pares de usuários tende a ser pequeno, assim como a centralidade de *closeness* e o coeficiente de clusterização dos vértices. Embora a comunidade EnglishLearning tenha um número menor de usuários, ela possui características semelhantes aos outros subreddits em relação às interações entre usuários, respeitadas as respectivas escalas.

Em relação à **QP2**, observamos que a distribuição da profundidade das árvores de discussão não possui cauda pesada. Enquanto os três primeiros níveis abaixo da raiz têm largura média maior que 2 para todos os subreddits, exceto English, a maioria tem largura média muito próxima de 1. Essas duas observações indicam que as interações dentre de uma thread costumam ser entre pares de usuários, e não de muitos usuários interagindo com uma mesma pessoa. Uma característica marcante do EnglishLearning é o alta proporção de *posts* sem resposta (46,45% dos *posts* analisados) em relação às outras comunidades (de 11,85% a 18,44%). É provável que isto esteja correlacionado ao baixo número de usuários na comunidade do inglês, embora não tenhamos investigado se há causalidade e, em que direção.

Em relação à **QP3**, baseado nas *tags* de nível de proficiência utilizadas no subreddit German, observamos que existem diferenças entre os textos de usuários com diferentes níveis de alemão quanto ao tamanho das palavras, uso de expressões idiomáticas, etc. Limitações desta análise incluem o fato da proficiência ser auto-declarada e de que muitos usuários não possuem *tags* de proficiência associadas a seus perfis. Como trabalhos futuros, pode-se investigar a evolução da proficiência dos usuários ao longo do tempo, à medida que eles interagem com a comunidade. Assim seria possível avaliar a eficácia da participação de comunidades online no aprendizado de idiomas.

Referências

Andalibi, N., Haimson, O. L., De Choudhury, M., and Forte, A. (2016). Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM.

Arnold, N. and Paulus, T. (2010). Using a social networking site for experiential learning: Appropriating, lurking, modeling and community building. *The Internet and higher education*, 13(4):188–196.

Cunha, T. O., Weber, I., Haddadi, H., and Pappa, G. L. (2016). The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 99–103. ACM.

De Choudhury, M. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1027–1036. ACM.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Lin, C.-H., Warschauer, M., and Blake, R. (2016). Language learning through social networks: Perceptions and reality.

Newman, M. (2011). Resource letter cs–1: Complex systems. *Am. J. Phys*, 79:800.

Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Silva, A. P. C., Meira Jr, W., and Beleigoli, A. M. R. (2017). Factors associated with weight change in online weight

management communities: A case study in the loseit reddit community. *Journal of Medical Internet Research*, 19(1).

Warschauer, M. and Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(2):57–71.

Zhao, Y. (1996). Language learning on the world wide web: Toward a framework of network based call. *Calico Journal*, pages 37–51.

Zourou, K. (2012). De l'attrait des médias sociaux pour l'apprentissage des langues–regard sur l'état de l'art. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 15(1).