Gender Bias in Turkish Word Embeddings: A Comprehensive Study of Syntax, Semantics and Morphology Across Domains

Duygu Altinok
Deepgram Research, USA
duygu.altinok@deepgram.com

Abstract

Gender bias in word representations has emerged as a prominent research area in recent years. While numerous studies have focused on measuring and addressing bias in English word embeddings, research on the Turkish language remains limited. This work aims to bridge this gap by conducting a comprehensive evaluation of gender bias in Turkish word embeddings, considering the dimensions of syntax, semantics, and morphology. We employ subword-based static word vectors trained on three distinct domains: web crawl, academical text, and medical text. Through the analysis of gender-associated words in each domain, we not only uncover gender bias but also gain insights into the unique characteristics of these domains. Additionally, we explore the influence of Turkish suffixes on word gender, providing a novel perspective on gender bias. Our findings reveal the pervasive nature of gender biases across various aspects of the Turkish language, including word frequency, semantics, parts-of-speech, and even the smallest linguistic unit - suffixes. Notably, we demonstrate that the majority of noun and verb lemmas, as well as adverbs and adjectives, exhibit masculine gendering in the general-purpose written language. This study is the first of its kind to offer a comprehensive examination of gender bias in the Turkish language.

1 Introduction

The rise of pretrained language models, such as BERT (Devlin et al., 2019), has significantly improved various natural language processing (NLP) tasks. However, these models are often trained on large amounts of web-based data, which can contain social stereotypes and biases that may be inherited by the models. This raises concerns as such biases can be perpetuated in downstream applications (Tal et al., 2022). The advent of large language models (LLMs) (Minaee et al., 2024) has further highlighted the importance of understanding and evaluating training data quality, including the presence of toxicity and gender bias (Zhao et al., 2024).

While previous studies have primarily focused on gender bias in embeddings, particularly in English, research on other languages has been limited to a few multilingual projects. For instance, (Prates et al., 2019)) evaluated gender bias in machine translation by translating gender-neutral languages using the Google Translate API, while (Lewis and Lupyan, 2019) examined gender stereotypes across 25 natural languages. However, languages other than English have received minimal attention in this research domain.

In the case of the Turkish language, the existing research is sparse. (Ciora et al., 2021) investigated overt and covert gender bias in machine translation models by examining gender-neutral Turkish and gendered English. (Caglidil et al., 2024) explored gender bias in Turkish transformer models. Despite the advancements in LLMs, several crucial research gaps related to the Turkish language still remain. Firstly, there is a lack of studies focusing on the fundamental form of word embeddings, namely pretrained word vectors. Secondly, we strongly believe that Turkish morphology warrants an extensive linguistic study that delves into the intrinsic nature of the language itself. This unique aspect of Turkish sets it apart from English and other welll-studied Western languages and adds an additional dimension to the research on gender bias.

To address these gaps, our work aims to fill the research void by conducting a comprehensive evaluation of gender bias in Turkish word embeddings, considering the dimensions of syntax, semantics, and morphology. We employ static embeddings, specifically Floret vectors, trained on three distinct domains: web data, academic data, and medical data. Through our analysis, we investigate the frequency of words associated with men and

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 203–218

women, examine the parts of speech associated with each gender, and explore the conceptual clusters of words associated with men and women. Additionally, we provide an in-depth exploration of the relationship between morphology and the gendering of words by dissecting the semantic aspects of suffixes.

Our main contributions are as follows:

- We conduct a comprehensive study on gender bias in Turkish word embeddings, which is the first of its kind.
- We consider syntax, semantics, and morphology dimensions in our work, across three distinct domains.
- We demonstrate that gender biases are prevalent across various aspects of the Turkish language, including word frequency, semantics, parts-of-speech, and even in the smallest unit of the language - suffixes. We show that in the general-purpose written language, the majority of noun and verb lemmas, as well as adverbs and adjectives, are gendered masculine. The majority of art, sports, and profession-related noun lemmas are also masculine, along with abstract nouns, body parts, electronic devices, clothing, and everyday object names.
- We also demonstrate that word morphology directly impacts the gender of word forms and can switch the gender of word forms. We research which suffixes have which gender impact on the word form.

This paper is organized as follows: we present our data and methodology, followed by domain-specific results related to the pretrained embeddings in each domain. The final section focuses on morphology. Our code and data are available in our Github[1] and Huggingface[2] repositories.

2 Methodology

In this section, we provide an overview of our methodology, including details about the datasets used, the choice of word vectors, and the process of training and calculating gender-related metrics.

2.1 Data

We utilized three distinct domains to train and evaluate our word vectors. The first domain is web crawl data, obtained from the mC4 part of the CulturaX dataset (Nguyen et al., 2023). This corpus, consisting of 76,432,893 documents, serves as a reflection of societal consciousness and is commonly utilized in various NLP tasks. To ensure data quality, we performed cleaning and preprocessing on the web crawl corpus, applying additional filters to enhance its overall reliability.

The second domain focuses on academic papers, where we expect minimal gender bias. We collected this data from various sources, including YÖK Açık Erişim [3] and Dergipark [4]. Both organizations, affiliated with the government, provide high-quality research papers and journals on their respective websites. We compiled abstracts from these sources, resulting in a total of 309,169 abstracts from YÖK Açık Erişim and 188,106 abstracts from Dergipark. Additionally, we obtained full article bodies solely from Dergipark, comprising 147,961 documents. The combined dataset from these sources is referred to as Academic Crawl, with a total of 645,236 documents.

The third domain focuses on the medical field and involves crawling research papers from Dergipark. We specifically selected journals with a medical focus, resulting in a corpus of 37,910 documents. Similar to the Academic Crawl corpus, the Medical Crawl corpus underwent cleaning and processing, including language filtering and the resolution of PDF-to-text errors.

Regarding the Medical Crawl corpus, we managed to eliminate PDF-to-text mistakes by implementing rules targeting single characters and missing vowels/consonants in between. Only a small portion of the data required removal.

In handling the Academical Crawl corpus, we faced a higher frequency of errors and articles with mistakes, presenting a more complex task. To address this, we conducted experiments and observed the effectiveness of the LLM Qwen2-7B (Bai et al., 2023) of recognizing Turkish at the character level. Utilizing a single NVIDIA A100 80GB GPU, we dedicated 108 hours to process 4.3GB of data using a zero-shot configuration.

For reference, the sizes of each corpus are summarized in Table 1. Each corpus is available in

[1] https://github.com/DuyguA/GeBNLP-2024-Gender-Bias-Turkish-Word-Embeddings

[2] https://huggingface.co/turkish-nlp-suite

[3] https://acikerisim.yok.gov.tr/acik-erisim

[4] https://dergipark.org.tr/en/

Dataset	Size	Words
mC4	172.7GB	20B
Academic Crawl	4.3GB	480M
Medical Crawl	178.6MB	20M

Table 1: Sizes of the datasets used in the study: Measured in UTF-8 bytes and number of words (in billions/millions).

their respective Huggingface repositories [5].

3 Training and evaluation of word embeddings

In our study, we considered the agglutinative nature of the Turkish language with its rich morphology. To effectively represent the complex word forms that can be generated through the addition of numerous inflectional and derivational suffixes, we chose to use Floret embeddings (Bojanowski et al., 2017), an extended version of fastText (Joulin et al., 2016) that incorporates Bloom embeddings (Grave et al., 2017). Floret combines word and subword information, allowing for more compact vector tables with enhanced representation of the morphological structure. Compared to traditional word vectors, Floret's subwords are up to 10 times smaller.

We utilized the Floret library code by spaCy (Honnibal and Montani, 2017) to train our word vectors. The training was conducted using the Continuous Bag-of-Words (CBOW) algorithm (Mikolov et al., 2013), with each word vector having a dimension of 300. For the subwords, we considered 2-grams to 5-grams. To reduce the size of the vocabulary, we used a compact vocabulary of 250,000 entries for the web crawl corpus (mC4) and the Academic Crawl corpus. For the Medical Crawl corpus, which has a smaller size, we used a vocabulary of 80,000. The choice of the subword window range $[2, 5]$ was determined heuristically, considering that the length of most common Turkish suffixes varies from 1 to 5.

To evaluate the quality and effectiveness of the produced embeddings, we compared them to the Floret vectors of the pretrained spaCy model tr_core_news_lg (Altinok, 2023) on morphology and syntax tasks. We initialized a spaCy model with our Floret vectors and then trained syntactic parser components, including the POS tagger, dependency parser, and morphologizer components, on the BOUN treebank (Türk et al., 2022). Testing was performed on the test division of this treebank. The results, shown in Table 2, indicate that our Floret vectors perform well. It is worth noting that the spaCy Turkish model used for comparison were trained approximately one year ago, while our vectors have been trained on a larger corpus (mC4) with additional vocabulary. The Academic Crawl and Medical Crawl datasets are comparatively smaller in size and have a more focused and limited vocabulary. As a consequence, the performance of the word vectors trained on these datasets may appear slightly inferior when compared to the larger web crawl corpus.

To assess the gender bias encoded in the trained embeddings, we employed the method introduced by (Bolukbasi et al., 2016). For each word, we calculated a gender bias score by computing the dot product between its vector and the vector representation of the concept of "woman" subtracted by the vector representation of "man." In our experiments, we used the Turkish translations of "woman" (*kadın*) and "man" (*erkek*). A positive score indicates a closer association with masculinity, while a negative score implies a stronger association with femininity. The magnitude of the score reflects the degree of bias, with higher absolute values indicating greater bias. A score of 0 indicates neutrality. Unlike many other studies, our approach is unsupervised, and we did not employ the Word Embedding Association Test (WEAT) scores.

4 Results and discussion

In this section, we thoroughly examine each data genre individually. We first train the Floret word vectors for each genre and analyze the gender distribution of the vocabulary. We then provide statistics on the vocabulary based on different syntactic categories. Additionally, we conduct unsupervised clustering separately for each gender using all the words and explore the topics associated with each gender, focusing specifically on the web domain due to the more diverse range of topics compared to academic and medical papers. Finally, we delve into the relationship between morphology and gender by investigating how word genders change based on suffixes.

[5] Each dataset exists with their original Turkish name in our Huggingface repository. We used English translations to reach a broader audience. Names of the datasets we used are *clean-mC4*, *Akademik-Makaleler*, *Akademik-Ozetler*, *Medikal-Ozetler* and *Medikal-Makaleler*, respectively.

Model	POS acc	Morph acc	Lemma acc.	DEP-UAS	DEP-LAS
tr_core_news_lg	0.90	0.89	0.82	0.73	0.63
tr_gender_web_lg	0.91	0.92	0.86	0.73	0.65
tr_academic_web_lg	0.75	0.79	0.72	0.65	0.61
tr_biomed_web_lg	0.70	0.75	0.70	0.60	0.60

Table 2: The table displays POS accuracy, morphological analysis accuracy, lemma accuracy, unlabelled attachment score for dependencies, and labelled attachment score for dependencies. Accuracy scores are calculated by the spaCy trainer using the test sets. In a spaCy model, each pipeline component assigns relevant attributes (POS tag, morphological analysis string, lemma, dependency tag, and head in the dependency tree) to tokens. Accuracy for POS tag, lemma, and morphology is determined by collecting the attributes for each token and comparing them to the ground truth list. The last two attributes are evaluated at the syntax tree level, assessing the structure of the dependency tree, correct head, and dependency arcs. UAS measures structure accuracy, while LAS additionally evaluates the accuracy of dependency labels on each arc.

4.1 mC4

In this section, we present the results of gender analysis conducted on the Floret word vectors trained on the mC4 corpus, which consists of 250K vocabulary words. Figure 1 displays the distribution of gender within the vocabulary words. The findings reveal a significant gender bias, providing empirical evidence for the existence of "masculine defaults" in these large text corpora (Cheryan and Markus, 2020). Specifically, 87% of the vocabulary words in our word vectors are associated with men, while only 13% are associated with women.

We further examine the distribution of gender bias in syntactic categories. Unfortunately, the situation remains disheartening as women are severely underrepresented in certain categories. Verbs, for instance, predominantly belong to the masculine category, with only a few "feminine" verbs such as "süslenmek" (to dress up), "kremlenmek" (to apply body lotion), "güzelleşmek" (to become beautiful), and "güzelleştirmek" (to make someone/something beautiful) falling into the feminine category. Figure 2 provides a visual representation of this distribution.

Nouns also exhibit a skewed gender representation, with the majority of feminine words being proper nouns, including female names in both Turkish and other languages (e.g., Fatma, Emine, Madeline, Anya, Donna, Minerva, Mary). Only a limited number of nouns are categorized as feminine, such as "tanrıça" (goddess), "kraliçe" (queen), "imparatoriçe" (empress), and nouns that are commonly associated with femininity in society, such as "güzellik" (beauty), "makyaj" (make-up), and "ev" (home), along with their derivations and inflections. On the other hand, the masculine category encompasses a wide range of nouns, including abstract nouns, body parts, clothing names, electronic devices, and everyday objects.

Profession names also display a gender bias, with feminine professions limited to nurse, midwife, nanny, gymnast, dancer, make-up artist, florist, fashion designer, model, actress, stylist, and hairdresser. All other professions, including academician, professor, doctor, engineer, architect, journalist, pharmacist, economist, embryologist, detective, carpenter, tailor, movie director, violinist, cellist, painter, as well as leadership positions such as governor, boss, director, CTO, and CEO, are categorized as masculine. Even prominent tech company names like Google, Facebook, Alibaba, Aselsan, Havelsan, Roketsan, and TAI are masculine. Additionally, sports names, including volleyball and tennis, predominantly fall into the masculine category, despite being commonly played by women.

Adverbs and adjectives also exhibit a similar bias, with only a few feminine adverbs and adjectives related to grace and beauty. The masculine category encompasses a wide range of adjectives and adverbs, including both positive and negative meanings. It is worth noting that negative meanings do not relate to a specific gender, while the masculine category includes both positive and negative aspects of the same word.

Pronouns, including personal, interrogative, definite, and indefinite pronouns, overwhelmingly belong to the masculine category. Out of 973 pronouns in the vocabulary, only 11 are categorized as feminine. For example, "Bunda" (bu/this locative), "kendime" (kendi/oneself dative), "kendimi" (kendi accusative), "kendinizi" (kendiniz/oneselves accusative), "kime" (kim/who dative), "neresi" (nere/where possessive), "nesi"

(ne/what possessive), "neyin" (ne/what genitive), "seninle" (sen/you instrumental case), "bunda" (bu locative), and "bunla" (bu instrumental case). Only one personal pronoun, "seninle" (with you), falls into the feminine category. The rest of the personal pronouns, along with their inflections and all other pronouns, are categorized as masculine. It is worth noting that suffixes can change the gender, which will be explored further in Section 5.

The gender bias in both frequency and syntax is evident in the results, with the vast majority of vocabulary words are being masculine. Furthermore, the lack of representation of women extends to all syntactic categories. Appendix A exhibits some words with syntax categories from this corpus for a more detailed view of the vocabulary.

Next, we delve into the clusters formed within the feminine and masculine word groups. We utilized the K-means clustering algorithm (Hartigan and Wong, 1979) and determined the optimal number of clusters using The Silhouette method (Rousseeuw, 1987). Eventually, we identified 6 distinct semantic groups within the feminine words and 11 within the masculine words, as depicted in Figure 3. The masculine clusters encompass various "serious societal matters", such as science and technology, arts and music, business, economy, and politics. In contrast, the feminine clusters associated with family, appearance, beauty, lifestyle, and domesticity reinforce cultural expectations for women to be "submissive" and "passive". Interestingly, even the arts, typically considered "a soft and feminine" domain in some cultures (Garlick, 2004), are predominantly represented by masculine clusters.

It is important to note that our dataset has been carefully filtered to exclude any obscene content or sexual profanity. Considering the tableau above, if such vocabulary words were present, they would most likely form a feminine cluster.

To comprehend the distribution, attributions, and findings discussed in the previous paragraphs, understanding the presence of patriarchy in Turkey is crucial. Despite formal rules promoting gender equality, patriarchal beliefs, values, and norms persist. Turkey's rankings in the Global Gender Gap Report by the World Economic Forum reflect this, with positions of 105th out of 115 countries in 2006, 130th out of 153 countries in 2020, and 127th out of 145 countries in 2024 (below the global average each year) [6].

While urbanization has led to advancements for women in Turkish society, it remains a patriarchal Muslim society where the family holds significant importance. Research on the strength of patriarchy focuses on factors such as religion, socio-economic class, and ethnicity. Empirical analyses, like those conducted by (Ozdemir-Sarigil and Sarigil, 2021), reveal the persistence of powerful and widespread patriarchal values and understandings in Turkish society. These values are influenced by both material and ideational factors. Notably, religiosity contributes to the reinforcement of patriarchal tendencies, and men exhibit significantly stronger patriarchal values compared to women. Additionally, patriarchal tendencies tend to increase with age, indicating generational differences in patriarchal values.

According to research by the Kadir Has University Women and Family Studies Research Center, Turkish women face numerous challenges, including violence, unemployment, lack of education, street harassment, family pressure, gender inequality, and social pressure [7]. Studies such as (Özcan et al., 2016) and (Guvenc et al., 2014) highlight the pervasive issue of domestic and intimate partner violence against women, often resulting in fatalities. Femicide, especially the killing of women by intimate partners, is also a significant concern (Cetin, 2015; Erükçü Akbaş and Karataş, 2024). The wage gap in the workplace is another concern, but the safety and protection of women's lives take precedence even before discussing economic disparities.

In the context of patriarchy, where men are considered leaders and women are expected to be submissive and passive, the adjectives "serious societal," "submissive," "passive," "soft and feminine" used in the previous paragraphs align. Women are often relegated to subordinate roles, leaving the task of shaping society to men. This perspective reinforces the findings of our research, which supports sociological work indicating that women face challenges in Turkey. Our study demonstrates how deeply ingrained patriarchy is within the culture and how it influences language as well.

[6] https://www3.weforum.org/docs/WEF_GGGR_2024.pdf

[7] https://gender.khas.edu.tr/en/survey-public-perceptions-gender-roles-and-status-women-turkey

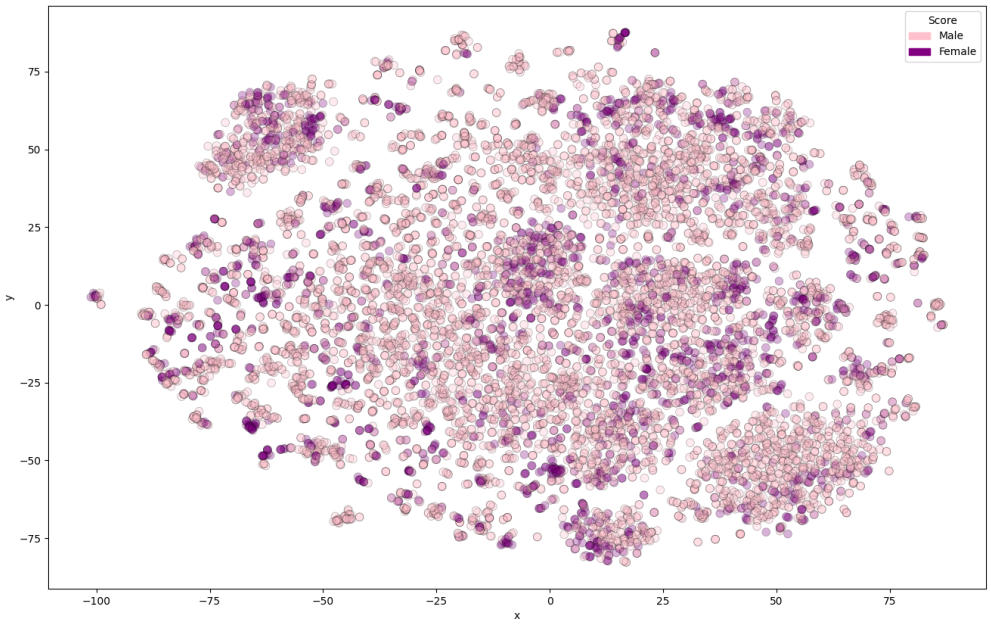

Figure 1: Visualization of gender associations in the vocabulary words represented by the 300-dimensional Floret embeddings. The vectors are reduced to 2 dimensions using the T-SNE algorithm (van der Maaten and Hinton, 2008). The visualization highlights that the online language space predominantly aligns with masculinity rather than femininity.

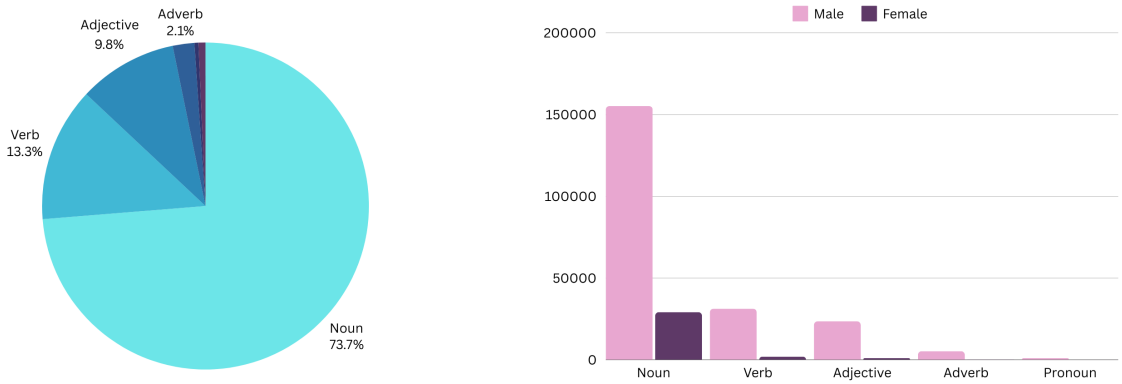

Figure 2: Distribution of syntax categories in the mC4 corpus, depicted as percentages (left). Gender distribution of vocabulary words within each syntax category (right).

4.2 Academic Crawl

In the following analysis, we examine word vectors trained on academic papers, with a vocabulary size of 250,000. We initially anticipated this domain to be relatively neutral; however, the distribution of gender associations turned out to be similar to the web domain. Approximately 12% of the words in the academic corpus are associated with the female gender, while the remaining 88% are associated with the masculinity. Figure 4 presents the corpus statistics and the distribution of syntax categories based on gender.

Unlike the web domain, there is a lesser presence of words related to traditionally "feminine" areas such as cosmetics and domestic topics. However, the number of health and science-related words has increased, resulting in a relatively stable count of adjectives and adverbs.

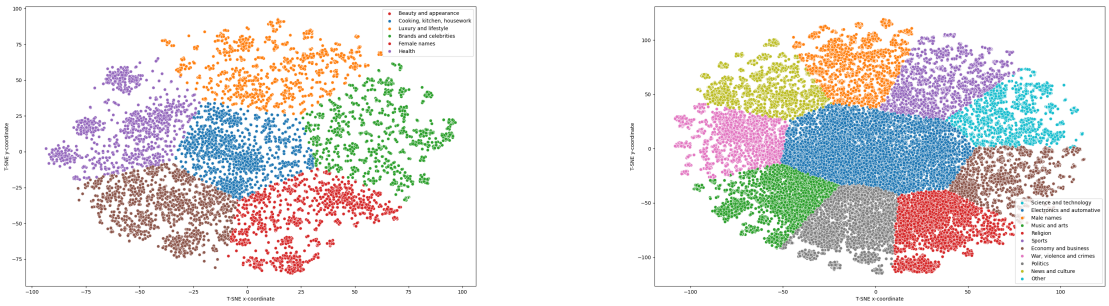

Figure 3: A visual comparison of semantic representations. (left) Femininity: Emphasizing traditional gender roles of homemaking, focusing on appearance, and limited involvement in professional life. (right) Masculinity: Freedom to explore various subjects and pursuits including fields of art, science, and professions.

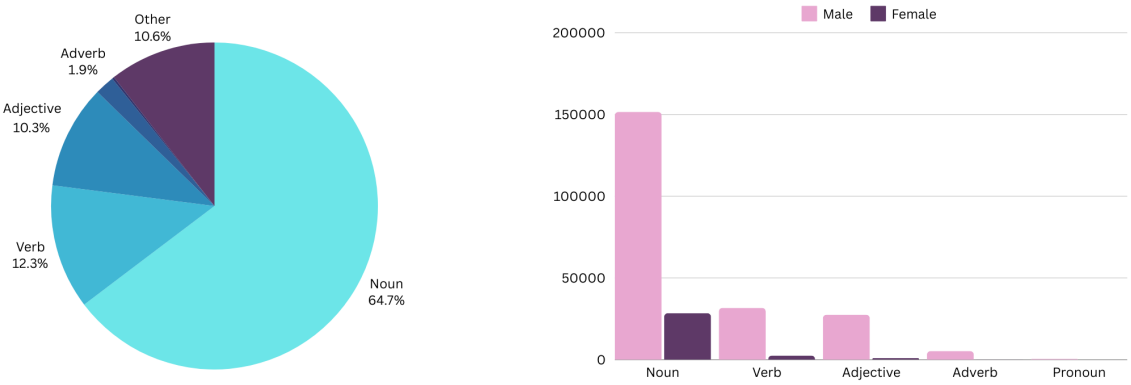

Figure 4: On the left, the distribution of syntax categories in the Academic Crawl dataset is presented, represented as percentages. On the right, the gender distribution of vocabulary words within each syntax category is depicted.

Moving on to verbs, there are notable differences in this domain. The majority of verbs are denominal verbs, which are formed by using a noun as a copula and transforming it into a verb, such as "olaydır" (olay + dır) and "değerleridir" (değer + leri + dir). It is expected to have a significant number of denominal verbs in this domain, as copular verbs are commonly used in formal and academic Turkish writing. Surprisingly, most of the feminine verbs are nominal verbs. The first group comprises nouns that are initially gendered as male but are then inflected with a copula and several other suffixes to become feminine verbs. The second group includes feminine nouns that undergo nominalization and become feminine verbs. We further analyze gender changes through inflection in the morphology section (Section 5). Masculine verbs, on the other hand, remain relatively consistent with those found in the web domain, representing normal and common verbs in the Turkish language.

Nouns present a distinct pattern. Around half of the nouns are associated with the female gender, including certain scientific terms, public health words, geographical names, and female names. The other half consists of nouns with masculine lemmas that become feminine nouns after undergoing suffixation. The addition of certain suffixes, like the plural suffix, can alter the gender of a masculine lemma. Since academic papers often contain numerous plural and group nouns, these nouns contribute to the overall count of feminine nouns, compensating for the absence of explicitly "feminine" words. The types of suffixes that affect gender are further discussed in Section 5.

As anticipated, the corpus contains various scientific terms such as "adsorpsiyon" (adsorption), "basınç" (pressure), "indüksiyon" (induction), "formülasyon" (formulation), "elastikiyet" (elasticity), "difüzyon" (diffusion), "fauna," "flora," as well as names of sciences like "kimya" (chemistry), "biyokimya" (biochemistry), "psikoloji" (psychology), and certain philosophical terms like

"metafizik" (metaphysics), "oryantalizm" (orientalism), "epistemoloji" (epistemology), and "popülizm" (populism). Among these scientific and philosophical terms, some are masculine (e.g., epistemology), while some are feminine (e.g., chemistry, pressure). Appendix B provides a sample of feminine scientific terms, which constitute a considerable portion of the corpus.

Overall, the scientific vocabulary, formal nature of academic writing, and the specific types of suffixation in formal written language contribute to the inclusion of feminine words in this domain, resulting in gender word counts similar to those in the web domain. For a comprehensive word list in the academic domain, please refer to Appendix B.

4.3 Medical Crawl

Due to its smaller size, we opted for an 80,000 vocabulary size for the word vectors in the Medical Crawl corpus. Figure 5 illustrates the corpus statistics and the distribution of syntax categories by gender, which closely resembles the academic domain (depicted in Figure 4), albeit with a slightly higher percentage of feminine nouns and adjectives. One might anticipate a more balanced gender distribution in this domain; however, the proportions of feminine and masculine words remain similar to those in the web domain, with 13% of words are feminine and 87% are masculine.

The increase in feminine adjectives is primarily linked to the health vocabulary. Adjectives such as "medikal" (medical), "jinekolojik" (gynecological), "klinik" (clinic), "dermatolojik" (dermatological), "kronik" (chronic), and "kardiyak" (cardiac) predominantly exhibit feminine associations and are frequently encountered in the medical domain. Masculine adjectives, on the other hand, consist mostly of common words in the language, such as "ritmik" (rhythmic), "mutlu" (happy), "ailevi" (domestic), "keçe" (felt), "yağlı" (oily), "radyoaktif" (radioactive), "kilolu" (overweight), "glutensiz" (gluten-free), "manyetik" (magnetic), and "güncel" (current). However, some medical domain adjectives also exhibit masculinity. Examples of such words can be found in Appendix C.

Moving on to nouns, as mentioned earlier, numerous medical terms are gendered as female, constituting a significant portion of feminine nouns. The remaining feminine nouns originate from inflected masculine words, similar to the patterns observed in the academic domain. Further explanation regarding this phenomenon can be found in the morphology section (Section 5), and a list of such words is provided in Appendix C. Masculine nouns mostly consist of common nouns used in written language.

Regarding verbs, masculine verbs primarily comprise common words in the language. Most of the feminine verbs are nominal verbs, similar to those in the academic domain, where nouns originally gendered as male are inflected with a copula to transform into verbs. A smaller portion of feminine verbs are medical nouns that have undergone inflection with a copula to become verbs. The percentage breakdown of the first and second group of nouns is 85% and 15%, respectively. Examples from both groups can be found in Appendix C.

Overall, the results align with those observed in the academic domain, with the medical domain exhibiting a greater presence of feminine words to some extent.

5 Morphology and gender

This section of our research is not specific to any particular domain; instead, we focus on exploring the role of morphology. Specifically, we investigate how certain types of suffixes influence the gender of words. We examine each type of suffix in detail within this subsection.

As mentioned in previous sections, our choice to use subword-enriched word vectors is motivated by a significant factor: we aim to generate representations for suffixes as well. In Turkish, a typical word is composed of morphemes, a lemma, and a list of suffixes, with each suffix carrying its own meaning. In this section, we demonstrate the semantic impact of suffixes on the gender dimension. We discuss various groups of suffixes, providing further explanations and examples for each group. For a comprehensive understanding of Turkish morphology and detailed explanations of these suffix groups, refer to (Karslake, 2021).

5.1 Inflectional suffixes

5.1.1 Nominal suffixes

Number. The plural suffix "-lAr" has the effect of changing the gender of some lemmas, predominantly from masculine to feminine. This could be attributed to femininity often being associated with a communal sense, family, and cooperation, resulting in plural nouns being mostly represented as feminine. However, in some cases, this suffix can transform feminine words into masculine

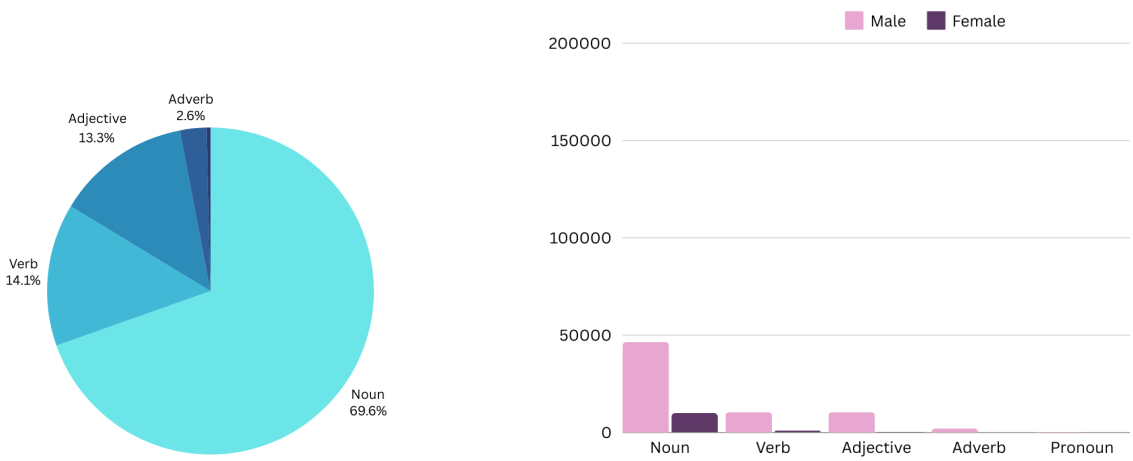

Figure 5: The percentages of syntax categories in the Medical Crawl dataset are illustrated on the left. On the right, the gender distribution of vocabulary words within each syntax category is displayed.

words, although this occurs less frequently. Figure 6 presents the statistics for the usage of the plural suffix in the three domains mentioned above.

Possession. With these groups of suffixes, we do not observe any significant gender-altering effects for singular persons. That is, possessive first, second, and third person singular suffixes ("-Im", "-In", "-(s)I") have minimal impact on the gender of the word. However, when it comes to plural person suffixes, there is a slight shift. The possessive first and third person plural suffixes ("-ImIz" and "-lArI"), for instance, can transform some masculine lemmas into feminine words (e.g., "ev+imiz" meaning "our house," "ev+leri" meaning "their house"). This may be attributed to the same communal sense observed with the plural suffix. On the other hand, the second person suffix ("-InIz") within this category does not have the same effect.

Case. We did not find any evidence suggesting that case suffixes are significantly related to gender. In very rare cases, these suffixes may change the gender of a word, but for the most part, they do not have an impact on gender.

ki. The suffix "-ki" serves two functions: when added to the genitive case of a noun, it forms a possessive pronoun (e.g., "kedi+nin+ki" meaning "one belonging to the cat"), and when added to the locative case of a noun, it creates an attributive adjectival phrase (e.g., "oda+da+ki vazo" meaning "the vase in the room"; "ön+ünüz+de+ki" meaning "the one in front of you"). In the first case, we did not find any instances of gender change. However, in the second case, "-dAki" does rarely alter the gender of both masculine and feminine lemmas. Nevertheless, we can conclude that "-ki" is not significantly related to the gender of a word.

5.1.2 Verbal suffixes

Voice. Causative, passive, reflexive, and reciprocal suffixes belong to this group. Except for the passive voice, we did not find any instances where verbs changed gender due to these suffixes. We believe that these suffixes have no significant effect on verb gender, except for the passive voice. The passive voice changes masculine verbs into feminine verbs, but not the other way around. We attribute this to the societal perception of females being associated with passive roles.

Negative marker. The negative marker "-mA" has an impact on verb gender. This suffix alone can change the gender of a verb (e.g., "üretmek(F)-üretmemek(M)," to produce and not to produce), and when combined with other suffixes, it can also alter the gender (e.g., "tanımak(M)-tanımamak(M)-tanıyacaksın(M)-tanımayacaksın(F)," to know, not to know, you'll know, you won't know). Most verb lemmas tend to maintain their gender when only the negative marker is added, accounting for approximately 5% of all verbs. However, when multiple suffixes are added, the possibilities become more varied, leading to new verb semantics.

Tense/aspect/modality. We did not come across any cases where verbs changed gender due to these suffixes.

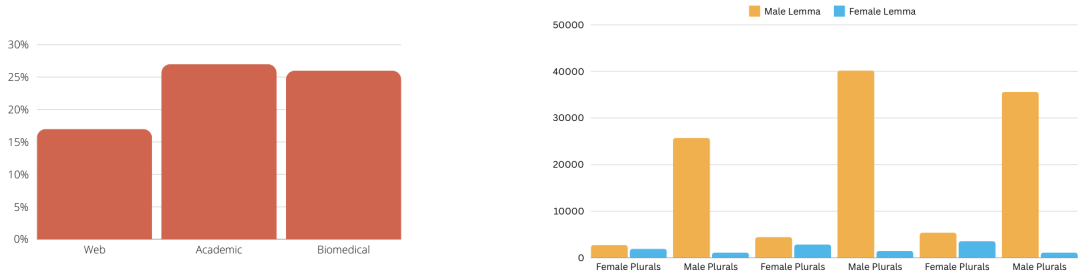

Figure 6: Left: Percentage of plural nouns compared to all nouns in each domain. Academic and medical domains have a higher proportion of plurals due to the text genre. Right: "Feminine/masculine plurals" indicates the number of plural feminine/masculine nouns. The columns represent the concentration of words with masculine lemma+plural suffix and feminine lemma+plural suffix. Around half of feminine plurals are derived from masculine lemmas, while masculine plurals are mostly inflected from masculine lemmas.

Personal markers. We did not find any instances where verbs changed gender due to these suffixes.

Copular markers. Copular markers present a different scenario. Based on the counts from the corpora of the three domains, the copula does not change gender. However, when inflecting a noun into a verb, several suffixes are typically added to the noun lemma, and some of them are gender-changing suffixes, such as the plural suffix (e.g., "göz(M)-gözler(F)-gözlerdir(F)"). As a result, the gender of the word changes. We observed that in the academic domain, many feminine verbs are formed in this manner, as masculine lemmas tend to shift towards feminine more frequently through suffixation.

Inflecting a verb from another verb is a completely different case; this situation can change the gender (e.g., "ağlamak(M)-ağlamış(M)-ağlamıştır(F)"). We already found out that most verb lemmas are masculine, and quite a few verbs in this category are found in academic and medical domains due to the formal written language, contributing to the count of feminine verbs.

Subordinate suffixes. In this section, we explore suffixes that convert verbs into nouns, adjectives, and adverbs. Participles (e.g., "yaratmak(M)-yaratan(M)," to create - creator) and converbs (e.g., "gitmek(M)-gider(M)-giderken(M)," to go, goes, while going) maintain the gender of the lemmas. However, when it comes to verbal nouns, the situation is slightly different. Some masculine verb lemmas, when suffixed to become a noun, transform into feminine nouns (e.g., "dokunmak-dokunma," to touch - the touch). We attribute this to actions being masculine, and when a word transitions from being a verb to being a noun, it loses the concept of action and becomes feminine.

5.1.3 Derivational suffixes

Derivational suffixes have the potential to shift the gender of words, although the reasons behind these shifts may not always be semantically clear. For example, in the triplet denge(M)-dengesiz(M)-dengesizlik(F) (balance-unbalanced-instability), the first word is the lemma and is masculine, while the derived forms are masculine and feminine.

Nominal->nominal derivation. In this category, masculine lemmas tend to shift towards feminine words more often than feminine lemmas. However, in rare cases, feminine lemmas can shift towards masculine words, as seen in the example sağlık (F)-sağlıkçı (M) (health-healthcare worker), where a concept transitions into a profession, which is typically associated with masculinity.

Nominal->verb derivation. Suffixes in this category typically shift feminine lemmas to masculine derived forms, as in the example sendika (F)-sendikalaşma (M) (labor union-unionization). This shift aligns with the observation that masculinity are more commonly associated with actions, as discussed in the previous section. Masculine noun

lemmas, on the other hand, do not change gender and become masculine verbs.

Verb->verb derivation. Suffixes in this category do not significantly change the gender. Since these suffixes do not alter the meaning of the verb substantially, the gender category remains the same.

Verb->nominal derivation. Most suffixes in this group do not change the gender. However, when a gender shift occurs, it predominantly affects masculine lemmas. Masculine verb lemmas become feminine nouns, as seen in the example toplanma(M)-toplantı(F) (to gather-a meeting). This can be explained by the association of actions with masculinity, so losing the aspect of being an action may imply becoming feminine.

We have provided examples of each gender-changing suffix in this section in Appendix D.

6 Conclusion

This paper examines the presence of gender bias in static word embeddings of the Turkish language. The findings indicate that gender biases are prevalent across various aspects, including word frequency, parts-of-speech, clustered concepts, word meaning dimensions, and even in the smallest units of the language, such as suffixes. Overall, the results reveal a pervasive association of words and concepts with men rather than women in Turkish pretrained embeddings. Furthermore, the study demonstrates how gender associations are differentiated based on parts-of-speech and clusters of concepts, with women being more associated with nouns and domestic content, while men are more associated with "serious matters." These findings raise concerns about the amplification of gender biases in AI and society through social, cultural, and digital mechanisms.

References

Duygu Altinok. 2023. A diverse set of freely available linguistic resources for Turkish. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13739–13750, Toronto, Canada. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Orhun Caglidil, Malte Ostendorff, and Georg Rehm. 2024. Investigating gender bias in turkish language models. *Preprint*, arXiv:2404.11726.

Ihsan Cetin. 2015. Defining recent femicide in modern turkey: Revolt killing. *Journal of International Women's Studies*, 16(2):346–360.

Sapna Cheryan and Hazel Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, 127.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Gamze Erükçü Akbaş and Kasım Karataş. 2024. Femicide in turkey: A document analysis of news from 2011 to 2019. *Journal of Social Service Research*, 50(1):54–72.

Steve Garlick. 2004. Distinctly feminine: On the relationship between men and art. *Berkeley Journal of Sociology*, 48:108–125.

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. 2017. Efficient softmax approximation for gpus. *Preprint*, arXiv:1609.04309.

Gulten Guvenc, Aygul Akyuz, and Sandra K Cesario. 2014. Intimate partner violence against women in turkey: A synthesis of the literature. *Journal of family violence*, 29:333–341.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *Preprint*, arXiv:1607.01759.

Celia Karslake. 2021. *Turkish: A Comprehensive Grammar*, 2nd edition. Routledge, New York.

Molly L Lewis and Gary Lupyan. 2019. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4:1021 – 1028.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Neslihan Keser Özcan, Sevil Günaydın, and Elif Tuğçe Çitil. 2016. Domestic violence against women in turkey: a systematic review and meta analysis. *Archives of psychiatric nursing*, 30(5):620–629.

Burcu Ozdemir-Sarigil and Zeki Sarigil. 2021. Who is patriarchal? the correlates of patriarchy in turkey. *South European Society and Politics*, 26(1):27–53.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate. *Preprint*, arXiv:1809.02208.

Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for Turkish dependency parsing: Introducing the BOUN Treebank and the BoAT annotation tool. *Language Resources and Evaluation*, 56(1):259–307.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *Preprint*, arXiv:2403.00277.

A Sample Words from the mC4 Dataset

Content Warning: The following word lists contains feminine and masculine words from the mC4 corpus, sorted by syntactic categories. The results might be triggering.

A.1 Female associated words

Noun aktris, allık, anası, annesini, aşk, bacım, bakım, bayanlar, bebek, blogger, çiçek, çiçekçi, çocuk, dansçı, dansöz, dekorasyon, Donna, ebeveyn, ev, evlilik, far, feminizm, fondöten, fotoğrafçı, güzellik, gül, hanımlar, hastane, hastanelerde, hemşire, hamilelik, halı, iletişim, İpekyol, kadınlarının, kadınlarımızı, kadınların, kadınlığını, Karısı, kızların, kleopatra, kraliçe, kuaför, kuaför, kıyafet, lale, Ludmila, magazin, makyaj, makyöj, manken, mankenlik, Mary, moda, modacı, mobilya, molina, mutfak, özbakım, oda, parfüm, podyum, Prada, Queen, rimel, ruj, sağlık, sivilce, stilist, süpürge, tanrıça, tasarımcı, temizlik, tedavi.

Verb aşılamak, beklemek, beslenme, bilmeliyiz, biliyorsun, boşanmak, boyamak, büyümek, büyütmek, çiçeklendi, çiçeklenmek, dayanışmak, doğan, doğurduğu, doğurduğum, doğuran, doğurmak, emzirmek, evlenmek, filizlenmek, flörtleşme, güzeldir, haberlerdir, hastalanma, iyileştirme, karşılaşılmaktadır, kadındır, karısıdır, oyalamak, ovalamak, oluşturuluyordu, silkeleme, silme, süpürme, temizleme, temizlerse, tanrıçaydı, yedim, yemedim, zayıfladım, zayıflama, zayıflayamadım.

Adjective beyazlı, döşemelik, erotik, esmer, güzel, kadife, kadifemsi, klinik, kozmetik, kumral, lezbiyen, mavili, narin, sarışın, simsiyah, sisli, süslü, yosma, zarif.

Adverb ağlarken, doğaçlama, doğal, dostça, güzelce, güzellikle, hamileyken, narince, soldukça, usulca, zarifçe, küstahça, yazın.

A.2 Male associated words

Noun adaylarını, ağabeyciğim, Ahmet, Akademi, akademisyen, aklınızdan, albümlerinde, Allah, analiz, araba, araştırmada, artezyen, Aselsan, asker,

atmosferde, baba, başmühendis, Belgeselin, belirti, belgeselini, beylerinden, bileklerinin, bilgisayar, bilimler, Bursaspor, cami, cemaatine, CEO, Charles, CTO, çavuş, çellist, dava, dedektif, demiryolu, deneyim, destan, devrim, doçent, doktor, doktrin, durum, domates, ekonomist, embriyolog, erkek, Eskimo, eşofman, eczacı, evresi, ezan, fermenstasyon, gazeteci, gençler, girişimci, gizliliğine, hareketlilik, Havelsan, heykelin, Hocamız, ihbar, insan, insanlarla, integral, intihar, isimlerinden, istatistik, ittifak, iyimserlikten, jeostrateji, kafatası, kahramanlar, kalabalıktan, Kanalının, Kardeşlik, kemancı, kesimlerinde, kapsülde, kısımlarda, kızılderilisi, kilise, konçerto, konsollar, kozmonot, liderinin, lig, macera, marangoz, masrafı, masrafları, mektuplarını, meslektaş, meyvesi, mimar, Mustafa, müfettişliğini, mühendis, oğlan, oğlum, oğulları, olay, onbaşı, Onur, operasyon, oyun, Peygamber, pilot, planör, profesör, proje, puma, rakiplerine, referandum, rehberlik, reklam, rektör, ressam, roket, Roketsan, sanatçı, savaşta, sembol, Sorumlulukların, soykırımın, subay, sultan, TAI, tarafımla, tatbikat, teknede, telefon, temas, terzi, toplantıya, Trevor, uydu, yönetici, yönetmen, yüzbaşı.

Verb açılmak, açmak, ayrılabilir, ayrılırlar, ayrılmak, buluşmak, buharlaşmak, dövüşmek, düşünmek, fokurdamak, gerilmez, görüşmek, haczetmek, hafifsemek, halletmek, hallolmak, hapsetmek, hapsolmak, hapşırmak, harcamak, hatırlamak, helalleşmek, hortlamak, hoşgörmek, hoşlaşmak, höykürmek, hükmetmek, hırpalamak, hıçkırmak, hışırdamak, ısırmak, ısıtmak, içmek, ikilemek, ilerlemek, iletmek, ilişmek, imrenmek, inmek, binmek, inanmak, incelmek, incelemek, incinmek, incitmek, indirgemek, ineklemek, inildemek, kalkmak, kapamak, kapışmak, kapanmak, kaydetmek, kaydolmak, kaçışmak, kaçmak, kırpmak, konuşmak, konuşturur, niyetlenmek, oturmak, pişmek, savsaklamak, sevinmek, sömürmek, sözleşmek, sözetmek, soruştur, süblimleş, tartışmak, tamamlamak, tamamlanmak, tekrarlamak, tıngırdamak, uçmak, uyanmak, uyandırmak, uyumak, uyutmak, yalvarmak, yakarmak, yerleşmek, yemek, yinelemek, zangırdamak.

Adjective absürt, ahlaklı, ahlaksız, akıllı, akılsız, alkolik, alternatif, amatör, aylak, ayrımsal, barbar, basamaklı, becerikli, belçikalı, bencil, bloke, bireysel, büyük, bütünsel, capcanlı, devrimsel, dik, dikili, dolandırıcı, doğulu, düşman, eğitimli, erişkin, erkeksi, evrendeki, faydalı, faydasız, gerçek, gerizekalı, günahsız, homojen, hırslı, hırssız, ilginç, insanlı, insansız, indirgeyici, ineklemek, indirgemek, inilmek, kanatlı, kanatsız, keyifli, keyifsiz, kesiklş, kişisel, kokusuz, kral, küçük, kurşunsuz, mükemmel, müşterek, müslüman, olgun, opsiyonel, pagan, paralel, pasif, periyodik, pratik, profesyonel, sahte, sahtekar, sağcı, salak, savaşçı, seçkin, stratejik, suçlu, suçsuz, tertemiz, teorik, yetişkin.

Adverb akıllıca, alçakça, amaçsızca, aniden, apaçık, aptalca, aksine, asıl, açıkça, baştanbaşa, beraberce, bilahare, büsbütün, büyükçe, dikkatlice, düpedüz, düşmanca, düşmanca, erkekçe, gençken, garipçe, hala, hariç, hızla, hızlı, hızlıca, henüz, hukuken, ileri, kasten, kazara, kısaca, kısaca, rahatça, saatlerce, saatlerce, saygısızca, sinsice, siyaseten, sırf, sürekli, tahminen, tamamen, tersine, uzaktan, yavaş, yavaşça, zaten.

B Sample words from Academic Crawl

B.1 Female associated words

The color green is used to visually highlight the masculine word lemmas within feminine words. This distinction specifically pertains to the lemmas within the set of feminine words.

Noun afetler, Avrupa, basınç, bilgi, Budizm, dermatoloji, dijitallik, dinamikler, düalizm, ekonomi, embriyoloji, Endokrinoloji, enformasyon, entegrasyon, etkileri, farmakoloji, felsefe, feministlere, feminizm, Fenomenoloji, fırsatları, fonksiyonu, fonksiyon, fonlar, formasyon, Gastroenteroloji, geleneği, gelişmeler, hastalık, Hititoloji, Hümanizma, ihtiyaçlar, iletişim, İmmünoloji, inanç, istatistik, Jinekoloji, Kadınlar, kadınlarımız, Kanji, Kardiyoloji, konular, komünizmi, kriptoloji, liberalizm, lirizm, literatür, Manihaizm, Meiji, mekanizması, metodolojilerle, Modernizm, motivasyonları, Müzikoloji, Natüralizm, oluşumu, ontolojisi, organizasyon, organizma, oteller, postmodernizm, psikoloji, Realizm, regresyon, rejimler, Rusya, Sağlık, simülasyon, sinizm, sorunların, Şamanizmi, tedavi, uygulamaları, üretim, ürün, veriler, volatilite, Wallis, Whitney, yapılarının.

Verb aynasıdır, bilinmektedir, bilmektedir, bilmektedirler, bilmekteyiz, bilmekteydiler, bilmişlerdir, bilmiştir, çiçeklenme, çiçektir, dengelenmesidir, dokumalardır, duygularıdır, fonksiyonudur, geleneğidir, gelmesidir, gerçekleşmemiştir, gıdalardır, görüştür, güzeldi, güzeldir, imgeleridir, inançlarıdır, kazandırmaktadır, kaybetmişlerdir,

kimlikleridir, koleksiyonudur, kuruluşlardır, kültürlerdir, kütüphanesidir, maliyetleridir, matrisidir, mekanıdır, mekanizmadır, mekanizmalardır, metinleridir, metodolojidir, oluşmuştur, oluşturabilmektedir, ortamlardır, sanatıdır, sanattır, sorunlardır, tedavidir, tedavisidir, teknikidir, ülkelerdir, ülkeleridir, üretmektedir, üretilmiştir, üretimdir, üretmektir, üretmekteydi, yapıdır, yapılardır, yapısıdır, yemeklerdir, yerleşimidir, yöntemdir, yöntemidir, yöntemlerdir, yöntemleridir.

Adjective anaerkil, bilgiişlemsel, cinsel, dokuma, elyazması, epidemiyolojik, estetik, farmakolojik, finansal, Kadınsı, logaritmik, lüks, magazinsel, medikal, rahmani, sismik.

Adverb güzellikle, gerçekleşmeyince, süslenip.

B.2 Male associated words

Noun anavatanlarına, arkeoloji, belediyeyi, dogmatizm, ekoloji, empatiye, emperyalizm, engellilerde, evrenselliğe, egzistansiyalizm, Eyaletler, feodalizm, geometrilerin, güvercinlerin, hademe, hiyerarşiden, hisse, hipofiz, imtiyazlara, insandan, işlevselliğinde, işkoluna, ısıyla, kademeleri, Katsayısının, kesitler, kiralardan, kirişlere, Koordinat, kozmogoni, kozmoloji, konstrüktivizm, levhalardan, liberalizm, macerasının, Nevrotiklik, nepotizm, nesnelliğe, oosit, oranlılık, organizma, oryantalizm, ozon, parlamenter, perakendecileri, prensip, profesörün, rasyonelizm, rasyonalizminin, şarkiyat, şarkiyatçı, sezonlarda, sevinçleri, sigortacılığının, sosyalizm, Stalinizm, Saltanatın, sembolizm, sembolizm, sübjektivizm, taşıtlardan, teknoloji, tipoloji, totemizm, uzuv, uzamlar, uzamlar, yağmurlarının, Örgütlerinin, öngörüsünün.

Verb alıkoydu, alınmıştır, almaktadır, anlatıyordu, bağlanıyordu, bilememektedirler, biliyordum, bitirdik, bulunmuştur, buluşmuştur, buyurdu, diyordum, duygu, duydular, durmayacağız, durdurulmuştur, durulmuştu, düzenlememiştir, düzenlenmemektedir, etmiştir, gerekmektedir, gereksinimleridir, getirmemelidir, getirmeyebilir, görmemişler, göstermektedir, karşılaşıldı, karşılanmadı, karşılamaktadır, kaybolmaktadır, kokuyordu, korkuyordu, kullanılmıştır, kurudu, olmamasıydı, olmuştur, olurdum, oluşturabilecektir, oluşturabilirler, oluşturulabilmektedir, oluşturuluyor, seçilmişti, seçmişti, söylüyordum, sağlanamadı, şiddetlendirmiştir, sürdürüyordu, tanımışlardır, tanıyor, ulaşamayız, ulaştırabilmektedir, üretiliyordu, üretiliyor, yapılandırılmaktadır, yapılmıştır, yaratırlar, yayınlanmasıdır.

Adjective ahlaklı, ahlaksız, alaylı, ayrımsız, Babasız, barbar, bohem, büyük, budaklı, çekirdekli, çetrefilli, cüretkar, dürüst, dörtlü, ekstrem, eksantrik, ereksel, evreli, geveze, heteroseksüel, ikili, ikincil, insanüstü, karbonik, karşılıklı, kuşaklı, kuşaksal, kurşunsuz, opsiyonel, sağcı, seçenekli, seçkin, sözleşmesel, sözleşmesiz, suçlu, taahhütlü, teist.

Adverb ahlaken, ahlaksızca, akılsızca, apaçık, apansız, baskılayarak, bilgilendirmeden, bilgilendirilerek, bilmeyip, borçlanarak, büsbütün, çabucak, çocukken, dokunup, donatıp, dürüstçe, ezbere, evrilerek, girerek, giydirerek, gençken, hoyratça, izletilerek, karşılanmadıkça, karşılaşılırken, reddedip, şekillendirip, silkinip, suçlanarak, tamamen, tercihen, usulca, ulaşılamazken, ulaşılınca, vekaleten, yıllarca, yaşlanınca.

C Sample words from Medical Crawl

C.1 Female associated words

The color green is used to visually highlight the masculine word lemmas within feminine words. This distinction specifically pertains to the lemmas within the set of feminine words.

Noun antiseptikler, antiserumlar, bağışları, başvurular, değerlerimizin, dinamiklerine, etkileri, getirileri, ihtiyaçları, ilaçları, imkanlarıyla, randevuları, ögeler, ödemeler, olgularıyla, olaylarıyla, bedenlerinin, pansumanları, kazançları, verileri, verilerin, yiyecekleri, yönetmeliklerin, akne, anafilaksi, anemi, anestezi, anesteziyoloji, ajitasyon, basınç, basıncı, biyokimya, bulantı, cilt, dahiliye, dejenerasyon, depresyon, difüzyon, drenaj, enfeksiyon, enfeksiyon, enjeksiyon, fizyoloji, fizyoterapi, hastaık, hastalığı, hastane, hemodiyaliz, hemşire, hemşirelik, hipertansiyon, hipoglisemi, infertilite, influenza, kardiyoloji, lupus, maliyet, mamografi, mesane, migren, miyokard, Obstetrik, poliklinik, psikiyatri, RNA, sedimantasyon, sintigrafisi, sistem, sağlık, sağlık, sağlığı, tedavi, tedavi, tedavisi, tıp, vitamin, vücut, yöntem, yöntemi.

Verb ağrısıydı, bilgilerdir, bilgilendirilir, bilinmektedir, bulunmamasıdır, cihazlarıdır, durumlardır, hastalığı, hastalıklardır, hastalıklarıdır, hastalıktır, hemşirelerdir, histerektomidir, infeksiyondur, inançlardır, kadındı, kadındır, kaynakıdır,

komplikasyonlarıdır, kuruluşlardır, kuruluştur, kültürüdür, lezyonlarıdır, olaydır, oluşmaktaydı, oluşmasıdır, oluşmuştur, oluşturabilmektedir, oluşturmaktadır, oluşturulamamıştır, oluşturulmaktadır, oluşturulmasıdır, oluşturur, radyoterapidir, sendromudur, şikayetlerdir, tedavidir, tedavisiydi, teknikidir, varlıkıdır, vitamindir, yöntemleridir, yöntemlerdir.

Adjective klinik, kadın, ana, medikal, dişi, finansal, jinekolojik, farmakolojik, salgın, cinsel, epidemiyolojik, kozmetik, estetik, Kardiyak, Kardiyovasküler, istatistiki, finanse, memeli, lösemili, kliniksel, lezbiyen.

Adverb evdeyken, konunca, bilgilendirilip, kasten, inince.

C.2 Male associated words

Noun ameliyat, antiserum, gösterge, idare, katılımcılar, tükenmişlik, atrofi, refleks, sinir, algılama, mülakat, laboratuvarlarına, bilim, sezaryen, işlevselliğe, projeksiyon, Histopatolojik, kongre, hidrolize, veziküller, genotip, önlem, adaptasyon, antidepresanlar, yetkinliklerini, Hesaplamalarda, kombinasyonları, semptomlardan, kalıtım, seratonin, bağlami radyolog, kontrendikasyonlar, aminoasit, rektum, hormon, sterilizasyonun, lezyon, lokalizasyonunda, rejenerasyon , anamnez, metastaz, olgularımızdan.

Verb ağrıdır, komplikasyonlardı, oluşturulabilmektedir, oluşturabilirler, değişikliklerdir, bildirmemiştir, yaratırlar, sağlanamadı, gerçekleştirebilmektedir, kimyasallardır, rastlamamışlardır, bilmiyor, oluşturulmalıdır, kaybetmiştir, karşılaşmamış, gerçekleşebilmektedir, kaybedilmiştir, karşılaştık, yitirmektedir, kümesidir, oluşturmalıdırü oluşturmuştur, oluşturmaktadır, olmaktadır, olmaktaydı, oluşturmaktaydı, oluşturabilir, oluşabilir, ulaşabilmektedir, ulaşamamıştır, ulaşmıştır, kaybedebilir, kazanabilir.

Adjective hasta, yüksek, önemli, anlamlı, düşük, ortalama, bağlı, cerrahi, farklı, istatistiksel, büyük, normal, ilişkili, genel, sosyal, aynı, farklı, etkili, pozitif, gerekli, uygun, nadir, sürekli, bilimsel, ekonomik, sınırlı, riskli, fonksiyonel, radyolojik, mümkün, kaynaklı, kardiyak, bütün, spesifik, yükseki toplumsal, dirençli, alternatif, dış.

Adverb olarak, olup, birlikte, erken, özellikle, sırasında, kullanılarak, karşı, alınarak, hızlı, bakımından, edilerek, giderek, yapılarak, uygulanarak, esnasında, yalnızca, takiben, ederek, kullanarak, aracılığıyla, değerlendirilerek, başlamadan, dayanarak, çabucak, yapmayıp, yapmazken, sorunsuzca, yemeden, akıllıca, puanlanırken, kullanmaktayken, gizlice, gerekmedikçe, geciktirilmeden.

D Sample surface forms of gender-altering suffixes

All the examples are derived from the word vectors that have been trained on the mC4 corpus.

D.1 Nominal suffixes

Plural suffix -lAr. dinamikler, gecekondular, geziler, görüşmeler, ihtiyaçlar, kitaplar, konular, müzeler, süreçler, surlar, tapınaklar, tesisler.

Possessive suffixes -ImIz and -lArI. akılları, aklımız, bilinçaltımız, bugünümüz, büromuz, cumhuriyetimiz, dergimiz, memleketimiz, midemiz, odaları, ormanlarımız, Programımız, şirketimiz, taleplerimiz, tanıtımımız, tulumları, yöneticimiz.

In the provided examples, the lemmas are exclusively masculine, while the resulting word forms are exclusively feminine. The red color is used to indicate the passive voice marker in the examples.

D.2 Verbal suffixes

Passive voice. gitmek-gidilmek, üretmek-üretilmek, yakaladı-yakalandı, yakalamak-yakalanmak, yakalayacaksın-yakalanacaksın.

In the provided examples, the lemmas are exclusively masculine, while the resulting word forms are exclusively feminine. The red color is used to indicate the passive voice marker in the examples.

Negative marker -mA. bilmek(F)-bilmek(M), tanırım(M)-tanıмam(F), tanıyacaksın(M)-tanımayacaksın(F), üretmek(F)-üretmemek(M).

The red color is used to indicate the negative marker in the examples.

Copular markers. ağlamak-ağlamış-ağlamıştır, bilmek-bilmiş-bilmişlerdir, gelmek-gelmiş-gelmiştir, gitmek-gitmiş-gitmişlerdir, kaybolmakta-kaybolmaktadır, oluşmuş-oluşmuştur.

In the provided examples, the lemmas are exclusively masculine, while the resulting word forms are exclusively feminine. Copular markers are indicated with the red color.

Subordinate suffixes. dokunmak(M)-dokunma(F), kapamak(M)-kapama(F), uymak(M)-uyma(F), dolamak(M)-dolama(F).

D.3 Derivational suffixes.

Nominal->nominal. denge(M)-dengesizlik(F), kimse(M)-kimsesizlik(F), nem(M)-nemlilik(F), şair(M)-şairlik(F).

Nominal->verb. fena(F)-fenalaşmak(M), flu(F)-flulaşmak(M), kadife(F)-kadifeleşmek(M)

Verb->nominal bulanmak(M)-bulantı(F), görmek(M)-görenek(F), toplanmak(M)-toplantı(F).