Project I: Analyzing language
Language as Data at Göttingen University
Moujan Mirjalili, Farhan Kayhan, Merle (Joris) Hellwig 25.11.2024

Deliverable 1

▶ Persian (Farsi)
Names of the language
Persian, also known as Farsi "فارسی" in its endonym form. The name Persian is used in English, while Farsi is the native term used by speakers of the language.

Number of speakers
Persian is spoken by over 120 million people as their first or second language. It's the official language of Iran, Afghanistan (as Dari), and Tajikistan (as Tajiki).

Preservation status
According to UNESCO's Atlas of the World's Languages in Danger, Persian is classified as Safe in terms of its vitality and endangerment status.

Geographic region
Persian is predominantly spoken in Iran, Afghanistan, and Tajikistan. There are also significant Persian-speaking communities in Uzbekistan and other regions with a Persianate history. The language has a sizable diaspora spread across various countries.

Language family
Persian belongs to the Indo-European language family, specifically the Iranian branch of the Indo-Iranian subdivision. It's part of the Western Iranian language group.

Grammar
Persian typically follows a Subject-Object-Verb (SOV) word order. It's considered a fusional language, though it has evolved to become more analytic over time. It uses an ezafe construction to indicate some relations between words. The World Atlas of Language Structures (WALS) provides additional insights:

- Persian has no grammatical gender and has lost most of its case system.
- It uses postpositions rather than prepositions.
- The language employs a nominative-accusative alignment in its case marking.
- Persian has a rich system of verbal person marking.

Persian has a relatively simple phonology with 23 consonants and 6 vowels. It's not a tonal language. The language has developed compound verbs, which are formed by combining two words, usually a verb and a noun.

Orthography
Persian is written using a modified version of the Arabic script. This writing system is an abjad, where consonants are written as full letters and vowels are typically represented by diacritical marks. The current form of the script for Persian was established after the Islamic conquest of Persia in the 7th century CE. The orthography is not strictly phonemic, as some sounds are represented by multiple letters.

▶ Spanish
Names of the language
The English name of the language is Spanish. However, in Spanish, different terms are used to refer to the language. The term Castilian (castellano) is employed in official contexts, such as in the Spanish Constitution and other formal occasions. Meanwhile, the Royal Spanish Academy (Real Academia Española) primarily uses the term Española.

Number of speakers
Spanish is the second most spoken native language (L1) in the world after Mandarin Chinese, with approximately 488 million native speakers, making up around 6% of the global population. As a second language (L2), Spanish is spoken by about 6.9% of the world's population.

Preservation status
The World Atlas of Languages categorizes Standard Spanish as potentially vulnerable. However, this classification seems inconsistent with the large number of speakers worldwide. Given the high number of native and second-language speakers, it would be more logical to consider Spanish a safe and thriving language.

Geographic region
Spanish is spoken in many countries worldwide. Below is a list of Spanish-speaking countries by continent:

- North and Central America: Costa Rica (de jure), El Salvador (de jure), Guatemala (de jure), Honduras (de jure), Mexico (de facto*), Nicaragua (de facto*), Panama (de jure)
- The Caribbean: Cuba (de jure), Dominican Republic (de jure)
- South America: Argentina (de facto*), Bolivia (de jure), Chile (de facto*), Colombia (de jure), Ecuador (de jure), Paraguay (de jure), Peru (de jure), Uruguay (de facto*), Venezuela (de jure)
- Europe: Spain
- Africa: Equatorial Guinea

This list does not include the large Spanish-speaking communities in non-Spanish-speaking countries, such as the United States. Notably, Spain is the only Spanish-speaking country in Europe, even though the language originated there. A key reason for the widespread use of Spanish globally is its dissemination through colonization.

Language family
Spanish belongs to the Indo-European language family and traces its origins to Vulgar Latin.

Grammar

Spanish typically follows a Subject-Verb-Object (SVO) word order. It is a fusional language, meaning it uses inflections to convey grammatical relationships. Spanish features 18 consonants and 5 vowels, giving it a balanced consonant-vowel ratio. Nouns do not have a case system, but pronouns distinguish nominative, accusative, and dative cases. A notable characteristic of Spanish is its verb conjugation system, where verbs change based on tense, aspect, mood, and the subject. This allows the pronoun to often be implied within the verb form itself.

Orthography

Spanish uses the Latin alphabet, with additional characters such as ñ and diacritical marks like acute accents. Its orthography is largely phonemic, meaning most letters consistently correspond to specific sounds. The current spelling conventions have been standardized primarily by the Royal Spanish Academy (Real Academia Española).

Deliverable 2

▶ Datasheet for Persian Dataset

Composition:

5. **What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?**
   The instances represent words, sentences, and metadata from Persian-language news sources.

6. **How many instances are there in total (of each type, if appropriate)?**
   There are 100,000 instances (sentences and their corresponding words) in total.

7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   It's a sample of instances from a larger set, as there are other sizes (10K, 30K, 300K, 1M, 3M) available for many languages, including Persian.

8. **What data does each instance consist of?**
   Instances vary by table:
   - Sentences: sentence text, sentence ID
   - Words: word, frequency, ID
   - Word co-occurrences: word pairs, significance of their co-occurrence (In co_s and co_n tables the co-occurrence of words, including frequency and word pairs with IDs is captured.)
   - Sources: URL, date of download, sentence ID

9. **Is there a label or target associated with each instance?**
   No explicit labels or targets are associated. The dataset focuses on language statistics rather than classifications.

10. **Is any information missing from individual instances?**
    No information appears to be missing from individual instances based on the provided format.

11. **Are relationships between individual instances made explicit (for example, user's movie ratings, social network links)?**
    Yes, relationships are captured in word co-occurrence tables (co_s and co_n), showing associations between words.

12. **Are there recommended data splits (for example, training, development/validation, testing)?**
    No recommended data splits are provided.

13. **Are there any errors, sources of noise, or redundancies in the dataset?**
    The dataset undergoes cleaning processes, but some noise may remain, such as non-words or fragmented sentences or even typos.

14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?**
    The dataset is self-contained, but it includes URLs to the original sources of the sentences.

15. **Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
    The dataset does not contain confidential data as it's compiled from publicly available web sources.

16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
    As the content is from various web sources, it may contain potentially offensive or sensitive material, though this is not the focus of the corpus.

17. **Does the dataset identify any sub-populations (for example, by age, gender)?**
    The dataset does not explicitly identify sub-populations.

18. **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?**
    No, the dataset does not contain identifiable personal data. additionally, individual identification is unlikely as the focus is on linguistic patterns rather than personal data.

19. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
    The dataset may contain sensitive topics as it's derived from various web sources, but it doesn't focus on or categorize such information.

20. **Any other comments?**
    No.

Collection Process:

21. **How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?**
    The data was directly observed from web sources, primarily news sites and randomly selected web pages.

22. **What mechanisms or procedures were used to collect the data? (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)**
    Data was collected using automated web crawling and text extraction software developed by the Leipzig Corpora Collection team.

23. **If the dataset is a sample from a larger set, what was the sampling strategy? (for example, deterministic, probabilistic with specific sampling probabilities)**
    The sampling strategy involves randomly selecting sentences from the crawled web pages to create a representative sample of the language.

24. **Who was involved in the data collection process and how were they compensated? (for example, students, coworkers, contractors) and how were they compensated (for example, how much were crowd workers paid)**
    The collection was likely conducted by researchers at Leipzig University, but specific details about individuals and compensation are not provided.

25. **Over what timeframe was the data collected?**
    The data for this specific corpus was collected in 2020, as indicated by the corpus name.

26. **Were any ethical review processes conducted? (for example, by an institutional review broad)**
    Ethical review processes are not explicitly mentioned in the provided documentation.

27. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**
    The data was obtained from third-party sources (websites) rather than directly from individuals.

28. **Were the individuals in question notified about the data collection?**
    Not applicable, as the data are from publicly available news content.

29. **Did the individuals in question consent to the collection and use of their data?**
    Not applicable; data are sourced from public news.

30. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
    Not applicable.

31. **Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted?**
    No specific impact analysis is mentioned in the provided documentation.

32. **Any other comments?**
    No.

Preprocessing/cleaning/labeling:

33. **Was any preprocessing/cleaning/labeling of the data done? (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
    Yes, several preprocessing and cleaning steps were performed on the data. According to Goldhahn et al. (2012), the processing pipeline consists of the following steps:

a. HTML-stripping
b. Language identification
c. Sentence boundary detection
d. Cleaning (removal of non-sentences, duplicates, etc.)
e. Sentence scrambling (p. 760).

Additionally, the authors note that "For each corpus, word frequencies are calculated and significant word co-occurrences are extracted" (p. 760), implying tokenization and statistical processing. These steps ensure the creation of a clean, monolingual corpus suitable for linguistic analysis.

34. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? (for example, to support unanticipated future uses)
The raw data is not included in the downloadable corpus, only the processed data.

35. Is the software used to preprocess/clean/label the instances available?
The specific software used for preprocessing is not publicly available, but it is developed and maintained by the Leipzig Corpora Collection team.

36. Any other comments?
No.

▶ Datasheet for Spanish Dataset

Composition:

6. What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?
The instances represent sentences from the Spanish Language, scraped from the Spanish Wikipedia.

7. How many instances are there in total (of each type, if appropriate)?
There are 30,000 sentences and 62,3694 words in total.

8. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
It's a random sample of 30000 sentences.

9. What data does each instance consist of?
Each instance consists of a sentence ID and a sentence text.

10. Is there a label or target associated with each instance?
There are no explicit labels or targets.

11. Is any information missing from individual instances?
It is unclear whether entire Wikipedia articles were scraped or if parts of the articles are missing in this subset.

12. **Are relationships between individual instances made explicit (for example, user's movie ratings, social network links)?**
    No.

13. **Are there recommended data splits (for example, training, development/validation, testing)?**
    No recommended data splits are provided.

14. **Are there any errors, sources of noise, or redundancies in the dataset?**
    Yes, the dataset contains parts of Wikipedia pages that are not part of the language data. For example, it includes symbols used in Wikipedia for language representation.

15. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?**
    The dataset is self-contained.

16. **Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
    The dataset does not contain confidential data as it is compiled from Wikipedia which is public.

17. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
    Since the data comes from Wikipedia, it's possible that the dataset includes some complex topics. However, as we do not speak Spanish, we are unable to assess this.

18. **Does the dataset identify any sub-populations (for example, by age, gender)?**
    The dataset does not explicitly identify sub-populations. However, as Wikipedia is created by a big community and there are different topics, it is possible that in this sub set there are article about sub-populations.

19. **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?**
    Yes, this is possible, as Wikipedia includes articles about individuals. However, all the information in this dataset is publicly available through Wikipedia.

20. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
    No, but as stated in Question 17, the dataset may still cover some of these topics in a more general sense.

21. **Any other comments?**
    No.

Collection Process:

22. **How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?**
    The data was directly from Wikipedia website.

23. **What mechanisms or procedures were used to collect the data? (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)**
    Data was collected using automated web crawling and text extraction software developed by the Leipzig Corpora Collection team.

24. **If the dataset is a sample from a larger set, what was the sampling strategy? (for example, deterministic, probabilistic with specific sampling probabilities)**
    The sampling strategy involves randomly selecting sentences from the crawled web pages to create a representative sample of the language.

25. **Who was involved in the data collection process and how were they compensated? (for example, students, coworkers, contractors) and how were they compensated (for example, how much were crowd workers paid)**
    The collection was likely conducted by researchers at Leipzig University, but specific details about individuals and compensation are not provided.

26. **Over what timeframe was the data collected?**
    The data for this specific corpus was collected in 2021.

27. **Were any ethical review processes conducted? (for example, by an institutional review broad)**
    Ethical review processes are not explicitly mentioned.

28. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**
    The data was obtained from a third-party source (Wikipedia).

29. **Were the individuals in question notified about the data collection?**
    Not applicable, as the data are from publicly available in Wikipedia website.

30. **Did the individuals in question consent to the collection and use of their data?**
    Not applicable; data are sourced from Wikipedia website.

31. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
    Not applicable.

32. Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted?
No specific impact analysis is mentioned in the provided documentation.

33. Any other comments?
No.

Preprocessing/cleaning/labeling:

34. Was any preprocessing/cleaning/labeling of the data done? (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
Yes, several preprocessing and cleaning steps were performed on the data. According to Goldhahn et al. (2012), the processing pipeline consists of the following steps:

a. HTML-stripping
b. Language identification
c. Sentence boundary detection
d. Cleaning (removal of non-sentences, duplicates, etc.)
e. Sentence scrambling (p. 760).

Additionally, the authors note that "For each corpus, word frequencies are calculated and significant word co-occurrences are extracted" (p. 760), implying tokenization and statistical processing. These steps ensure the creation of a clean, monolingual corpus suitable for linguistic analysis.

35. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? (for example, to support unanticipated future uses)
The raw data is not included in the downloadable corpus, only the processed data.

36. Is the software used to preprocess/clean/label the instances available?
The specific software used for preprocessing is not publicly available, but it is developed and maintained by the Leipzig Corpora Collection team.

37. Any other comments?
No.

Deliverable 3

- **Number of word forms (tokens)**

Spanish: 623694
Persian: 2484412

- **Number of distinct word forms (types)**

Spanish: 79504 - 12.7% of all words in the dataset (Unless otherwise specified, the following will apply to the entire deliverable.)
Persian: 165300 - 6.7%

- **Sentence length in characters and words**

Spanish: 20.79 Words per Sentences 125.19 Chars per Sentences
Persian: 24.84 Words per Sentences 128.54 Chars per Sentences

- **Word form length**

Spanish:4.0252
Persian:4.0437

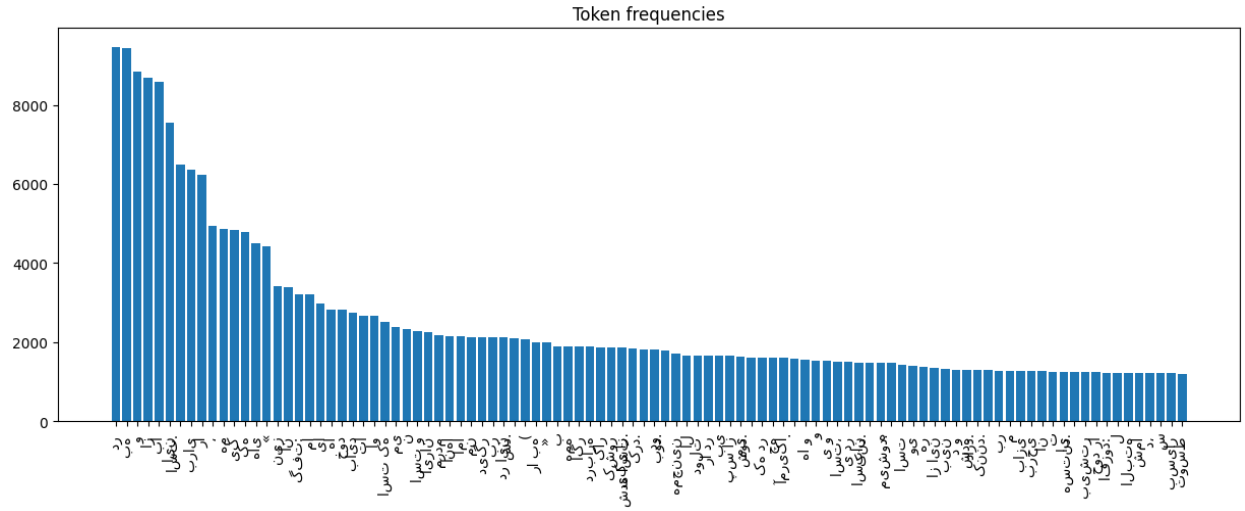
- **Word forms with frequency=1 (hapax legomena).**

Spanish:954 - 0.1529596%
Persian:152 - 0.0061181%

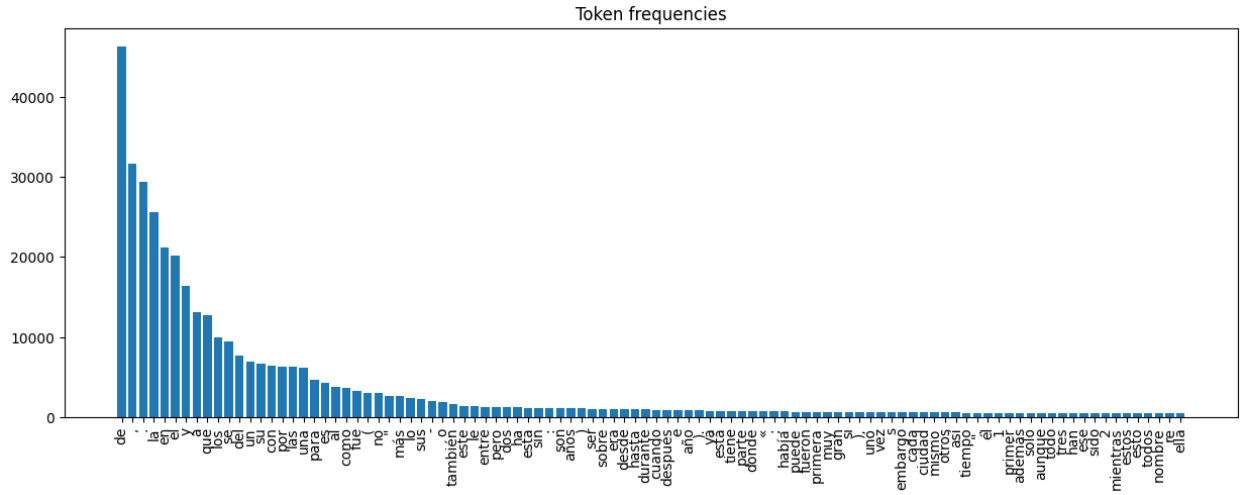
- **Most frequent words**

Spanish: de, la, en
Persian:در، به، و

- **Distribution of most frequent words**
  - **Persian**



- **Spanish**



- **Most frequent bigrams, trigrams, etc.**

	Spanish	**Persian**
Bigrams	'de', 'la' 'en','el' 'en', 'la'	'ادامه', 'داد' 'همشهری', 'نلاین' 'ویروس', 'کرونا'
Trigrams	'uno', 'de', 'los' 'por', 'lo', 'que' 'de', 'la', 'ciudad'	'گزارش', 'همشهری', 'نلاین' 'همشهری', 'نلاین', 'نقل' 'وی', 'ادامه', 'داد'
4-grams	‘a', 'lo', 'largo', 'de' 'por', 'primera', 'vez', 'en' 'la', 'mayoría', 'de', 'los'	'گزارش', 'همشهری', 'نلاین', 'نقل' 'همشهری', 'نلاین', 'نقل', 'ایسنا' 'ریس', 'دانشگاه', 'علوم', 'پزشکی'

- **Interrogative pronouns**

Spanish: 258 - 0.0413664%
Persian: 6758 - 0.2720161%

With Interrogative pronouns there is a similar case to make about the different type of datasets that were used. It’s is not possible to say wherever or not the difference is caused by this. However, we still choose this class as the other classes would have not given more information than this class.

- **Observations:**

Upon comparing the linguistic characteristics of Spanish and Persian, some interesting observations emerge. Notably, the languages show surprising similarities despite being from different linguistic families (Romance and Indo-Iranian), and using distinct writing systems (Latin script for Spanish, and Perso-Arabic script for Persian). Further details are outlined below:

- **Wordform Length:**
  Both Spanish (average wordform length of 4.03 characters) and Persian (4.04 characters) show nearly identical averages for word length, which is intriguing given the different writing systems.

- **Sentence Length:**
  The average sentence length in both Spanish and Persian is also very close, despite differences in grammar and syntax. This similarity suggests that sentence structure across these languages might share certain functional characteristics, possibly driven by the fact that both languages tend to favor subject-verb-object (SVO) word order in simple declarative sentences.

- **Number of Distinct Words:**
  A significant difference emerges when we look at the number of distinct words. Spanish has a notably higher number of unique words than Persian. However, this difference likely stems from the source and genre of the datasets rather than an inherent difference between the languages themselves. For instance, if the Spanish dataset comes from a source like Wikipedia, which tends to include a wide range of specialized vocabulary (scientific terms, technical jargon, etc.), it would naturally have a larger lexicon. In contrast, a news corpus, which might form the basis of the Persian dataset, typically uses a more standardized and less varied vocabulary, leading to fewer distinct words. Thus, the difference in distinct word count likely reflects the nature of the source material rather than linguistic characteristics of Spanish and Persian.

- **Hapax Legomena (Words Occurring Only Once):**
  The gap in hapax legomena (words that appear only once) between Spanish and Persian is even more pronounced. A larger number of hapax legomena often indicates a dataset that contains a broader range of specialized or rare terms, which is typically the case with sources like Wikipedia. Since Wikipedia articles are written in a more varied and diverse style, they tend to include more unique words. On the other hand, a news corpus usually contains more repetitive and standardized vocabulary, with fewer hapax legomena. Therefore, the larger number of hapax legomena in the Spanish dataset is likely a consequence of the type of source material used, rather than a true difference in the linguistic structures of the two languages.

Deliverable 4

1. Normalisation and Pre-tokenisation

- **Normalisation:**

  The Sequence of normalisers includes StripAccents and Lowercase. This approach ensures consistency in text processing by removing diacritics and converting all text to lowercase. Accents may not be consistently used in Spanish corpora, so stripping accents is a logical choice for robust tokenization. Spanish uses accents (á, é, í, ó, ú, ü) consistently for word differentiation (e.g., tú vs. tu). In tokenization for certain tasks, accents might be stripped using StripAccents, but this depends on whether morphological details are crucial. Lowercasing is helpful to unify case variations (Habla vs. habla).

  However, in persian accents are not used and there is no such thing as lower/upper case. We could not find anything else for normalizng so we used the same code as spanish for persian as well.

- **Pre-tokenisation:**

  A Whitespace pre-tokeniser is explicitly not applied for Persian because of the presence of multi-word expressions that might be semantically or morphologically linked, such as compound words and agglutinative constructs. Whitespace splitting is effective for Spanish, given its word boundaries are typically clear and consistent.

2. Key Observations from the Data

▶ **Persian:**

Frequent Words: On average, frequent words require ~2.26 tokens, indicating that even common words are sometimes split into subword units. For example, books might be split like this. (ketabha into ketab+ha)

کتاب‌ها ← کتاب + ها

Infrequent Words: These average ~2.72 tokens, reflecting increased segmentation for rarer words or morphologically complex terms. For example, petrochemicals might segment like this. (petroshimiha into petro+shimi+ha)

پتروشیمی‌ها ← پترو + شیمی + ها

Unseen Words:Unseen words are slightly less segmented than infrequent ones (~2.64 tokens). This could be because they contain subword fragments already common in the vocabulary.

▶ **Spanish:**

Frequent Words: The average token count is ~1.24, indicating that even common words occasionally undergo some level of segmentation. For example, articles (el, la), prepositions (de, en), and conjunctions (y, que) are likely tokenized as single units, but longer frequent words (e.g., rápidamente) might split into subwords like rápida + mente.

Infrequent Words: These average ~1.88 tokens, reflecting moderate segmentation for less common words. For example, technical or rarely used words like autopistas may segment as auto + pistas.

Unseen Words: Unseen words require ~2.58 tokens on average, showing substantial segmentation for out-of-vocabulary terms. For example, compound or domain-specific terms (e.g., electrodomésticos) might split into electro + domésticos.

3. **Segmentation Patterns**

- **Persian:**

Even frequent words often require subword tokenization, reflecting the language's morphological complexity (e.g., clitics, suffixes, compound words).

Tokenizers for Persian must work harder to balance between capturing morphological units and preserving meaning.

Frequent words average ~2.26 tokens, much higher than Spanish (~1.24). This suggests Persian's script and morphology create greater segmentation challenges, even for common words.

Unseen words (~2.64 tokens) are segmented slightly more than Spanish (~2.58 tokens), but the difference is minimal, indicating both languages struggle with unknown vocabulary.

- **Spanish:**

Frequent and infrequent words are segmented far less than Persian, reflecting the language's phonetic regularity and clear boundaries.

However, unseen words in Spanish are more segmented than frequent or infrequent ones, indicating increased complexity for out-of-vocabulary terms.

4. **Impact of Morphology and Orthography:**

- **Persian:**

The script's lack of explicit word boundaries (e.g., clitics like - م - my) and absence of short vowels contribute to higher segmentation counts.

Persian has a richer morphological system with affixes (e.g., کتاب‌هایم - my books) and clitics (e.g., کتابم - my book), leading to more frequent splitting into subword tokens.

Infrequent and unseen words often introduce more subwords because they combine uncommon fragments.

- **Spanish:**

Spanish's phonetic and consistent orthography allows the tokenizer to recognize complete words more easily, especially for frequent and infrequent words.

Spanish relies on suffixation (e.g., rápidamente, caminando) and compound words, which are generally less segmented but still show some inconsistency for infrequent/unseen terms.

For unseen words, derivational suffixes or compound words (e.g., automóviles) are often split, but the tokenizer performs better overall due to simpler word boundaries.

5. Practical Implications

Persian:

Efficiency Issues: High segmentation even for frequent words may slow down downstream tasks like translation or text generation.

Semantic Loss: Over-segmentation of infrequent and unseen words risks breaking key morphological units, such as clitics and plural markers.

Recommendations: Increase vocabulary size or use hybrid morphological and subword tokenization methods.

Spanish:

Balanced Performance: While frequent and infrequent words are well-handled, the tokenizer struggles more with unseen words.

Impact on Tasks: Machine translation may see issues with rare or compound terms. Named Entity Recognition (NER) might mislabel unseen entities due to token splitting.

Recommendations: Fine-tune the tokenizer for domain-specific corpora to reduce over-segmentation for unseen terms.

6. Broader Comparisons

Language	Frequent Avg. Tokens	Infrequent Avg. Tokens	Unseen Avg. Tokens
Persian	2.26	2.72	2.64
Spanish	1.24	1.88	2.58

Persian consistently has higher token averages, reflecting its complex script and morphology. Spanish shows smoother segmentation for frequent/infrequent words, but unseen words remain challenging for both languages.

7. Observations

When applying a subword tokenizer trained using Huggingface's tokenizers library, the segmentation results revealed notable trends and challenges, especially in languages with rich morphology like Persian and Spanish. The tokenizer's performance varied depending on the input structure, and its consistency depended on the training process and preprocessing decisions. Below, we provide examples from both languages and analyze the observed patterns:

- **Persian Corpus Analysis**

  For Persian, we trained the tokenizer on a dataset representative of the Hamshahri Online corpus, using Byte Pair Encoding (BPE) to address the language's morphologically complex nature. Pre-tokenization involved splitting based on whitespace and punctuation, without normalization of diacritics, as these are rarely used in Persian text.

Examples of Encodings:

[‘کلاه’, ’ت ’, ’را ’, ’پس ’, ‘می’, ‘دهم’, ‘.’]

[‘کل’, ’اهم ’, ’را ’, ’پس ’, ‘بده’]

[‘کلاه’, ’ش را ’, ’پس ’, ‘داد’, ‘.’]

[‘کلاه’, ’شان را ’, ’پس ’, ‘میده’, ‘یم’, ‘.’]

[‘کلاه’, ’تان را ’, ’پس ’, ‘بده’, ‘ید’, ‘.’]

[‘کلاه’, ’مان را ’, ’پس ’, ‘داد’, ‘ند’, ‘.’]

Correct Splits: In cases like “کلاهم” (my hat), the tokenizer correctly split the word into کلاه + م, aligning with morphological intuition. This success indicates the tokenizer’s ability to segment possessive markers accurately when the patterns were sufficiently represented in the training data.

Incorrect Splits: The tokenizer incorrectly segmented کل + اهم, where کلmeans “all” and اهم has no meaningful interpretation in this context. This reflects a limitation in distinguishing between compound words and concatenated morphemes.

Consistency in Affixes: Possessive markers (e.g., “مان“ ,”شان“ ,”م“) were segmented consistently when attached to frequent base forms (e.g., “کلاه”). However, rare or unseen combinations sometimes led to errors, highlighting the impact of token frequency on segmentation reliability.

▶ **Spanish Corpus Analysis**

For Spanish, the tokenizer was similarly trained using BPE, with pre-tokenization splitting contractions like del into de + el and preserving diacritics for semantic clarity.

Examples of Encodings:

[‘yo’, ‘me’, ‘du’, ‘cho’]

[‘tú’, ‘te’, ‘du’, ‘chas’]

[‘ella’, ‘se’, ‘du’, ‘cha’]

[‘nosotros’, ‘nos’, ‘du’, ‘cha’, ‘mos’]

[‘vos’, ‘otros’, ‘os’, ‘du’, ‘chá’, ‘is’]

[‘ellas’, ‘se’, ‘du’, ‘chan’]

Correct Splits: The tokenizer successfully segmented reflexive verbs like “ducharse” (to shower) into their base components (e.g., du + cha). Reflexive pronouns (e.g., “me,” “te,” “se”) were handled consistently.

Inconsistent Handling of Rare Forms: Less common forms, such as the Vosotros conjugation duchá + is, sometimes had inconsistent splits compared to more frequent conjugations like du + chan.

Regular Morphological Patterns: Verb endings (-ar, -mos, -is) were identified and segmented as expected, aiding the understanding of grammatical structure.

In conclusion, while the tokenizer demonstrated robust performance in frequent morphological constructions, inconsistencies with rare or complex forms underscored the need for improved training data and methodology. These findings suggest that tailored preprocessing and an enriched corpus are essential for achieving reliable subword segmentation in morphologically rich languages like Persian and Spanish.

Deliverable 5

▶ **Persian Overall Scores**:

- **Unlabeled Attachment Score (dep_uas): 83.59%**

This measures the percentage of correct head-dependent relationships, ignoring the labels. A high UAS indicates the model is good at identifying syntactic relationships, even if the labels are not always correct.

- **Labeled Attachment Score (dep_las): 78.78%**

This measures the percentage of correct head-dependent relationships where both the structure and dependency labels are correct. A lower LAS compared to UAS shows that while the structure is well-predicted, some labels are incorrect.

▶ **Spanish Overall Scores:**

- **Unlabeled Attachment Score (dep_uas): 82.92%**

This measures the percentage of correct head-dependent relationships, ignoring the labels. A high UAS indicates the model accurately identifies syntactic relationships, even when labels are not entirely correct.

- **Labeled Attachment Score (dep_las): 78.91%**

This measures the percentage of correct head-dependent relationships with both the structure and dependency labels correct. A slightly lower LAS compared to UAS reflects some label prediction errors.

Scores:

In the tables below, the macro-Averages treat each dependency type equally and have equal weight for each dependency type. micro-Averages are weighted by the number of true positives, false positives, and false negatives across all types and are weighted by total counts across types.

score	**Labeled Attachment Score (LAS)**			
Language	**Persian**		**Spanish**	
	Macro-Averages	Micro-Averages	Macro-Averages	Micro-Averages
Precision	0.7162	1.1071	0.705	1.047
Recall	0.6317	1.0669	0.652	1.040
F1-Score	0.6511	1.0866	0.674	1.044

score	Unlabeled Attachment Score (UAS)			
Language	**Persian**		**Spanish**	
	Macro-Averages	Micro-Averages	Macro-Averages	Micro-Averages
Precision	0.779	1.104	0.768	1.047
Recall	0.690	1.063	0.717	1.032
F1-Score	0.713	1.083	0.742	1.039

score	Label Accuracy Score (LS)			
Language	**Persian**		**Spanish**	
	Macro-Averages	Micro-Averages	Macro-Averages	Micro-Averages
Precision	0.645	1.107	0.634	1.047
Recall	0.569	1.067	0.587	1.040
F1-Score	0.586	1.087	0.607	1.044

Observations:

- **Corpus Source**
  Upon deeper analysis of the corpus, including the examination of the most frequent words, bigrams, and trigrams, it became evident that the data was not sourced from Wikipedia, as initially assumed. Instead, the recurring patterns in word combinations strongly suggested that the corpus was likely crawled from the Iranian news agency website, Hamshahri Online (همشهری آنلاین). This conclusion was reinforced by the high frequency of phrases directly associated with news reporting and journalistic terminology.

  For instance, some of the most common bigrams included (‘داد‘ ,’ادامه’) (continued saying), (‘کرونا‘ ,’ویروس’) (Coronavirus), and (‘همشهری‘ ,’گزارش’) (Hamshahri reported). Similarly, among the trigrams, phrases such as (‘نلاین‘ ,’همشهری‘ ,’گزارش’) (Hamshahri Online reported) and (‘ ,’وی داد‘ ,’ادامه‘’) (He/She continued saying) were prominent. The quadgrams further cemented this observation, with phrases like (‘نقل‘ ,’نلاین‘ ,’همشهری‘ ,’گزارش’) (Hamshahri Online reported quoting) and (‘کرونا‘ ,’مقابله‘ ,’ملی‘ ,’ستاد’) (National Coronavirus Task Force) pointing to specific news-oriented content.

- **Degrees of ccomp and acl-relcl**
  Analyzing the average degree of sentence components revealed that ccomp (clausal complements) and acl:relcl (relative clause modifiers) exhibited high degrees. This finding can be attributed to the nature of the corpus, which appears to contain a significant amount of content from a news website. Journalistic writing often involves reporting indirect speech and providing elaborate details through subordinate and relative clauses. For example, a sentence like "او گفت که وضعیت اقتصادی بهتر شده است" (He said that the economic situation has improved) illustrates how ccomp structures dominate such text. Similarly, acl:relcl usage is prominent in descriptive passages, contributing to the high degree counts observed. These patterns underscore the influence of corpus genre on dependency parsing metrics. Further details are outlined below:

  - **High Degree of ccomp**
    ROOT typically connects the main clause elements (subject, object, clausal complement). The ccomp dependency often captures complex subordinate clauses introduced by که (ke, "that"), which are integral to Persian sentence construction. In many cases, ccomp includes a full clause with its own subject, verb, and modifiers. This makes ccomp nodes inherently rich in dependents, contributing to their high average degree. Here is an example:

    او گفت که من باید این گزارش را تا فردا ارائه کنم.
    (He said that I must submit this report by tomorrow.)

    Dependency Structure:

    ROOT: گفت (said) governs two dependents:

    Subject: او (he).

    ccomp:

    که من باید این گزارش را تا فردا ارائه کنم (that I must submit this report by tomorrow).

    Degree of ROOT: 2.

    Degree of ccomp: 5, including:

    Subject: من (I). - Auxiliary verb: باید (must). - Object: این گزارش (this report). - Adverbial modifier: تا فردا (by tomorrow). - Main verb: ارائه کنم (submit).

    This example illustrates how ccomp often carries more complexity than the ROOT.

    As mentioned above, In journalistic or formal texts, ccomp structures are heavily utilized for indirect speech and detailed explanations. These subordinate clauses often have complex internal structures that outpace the simplicity of the ROOT node, which usually governs only high-level elements.

  - **High Degree of acl:relcl**
    The acl:relcl dependency, which corresponds to relative clauses, is another source of complexity in Persian syntax. Persian relative clauses are typically introduced by که (ke,

“that/who/which), similar to English, and are used to modify nouns by adding descriptive or explanatory content.

Relative clauses often encapsulate a full subordinate clause, including a subject, verb, and modifiers. For example:

کتابی که دیروز خریدم هنوز نخوانده‌ام.

(The book that I bought yesterday, I haven’t read yet.)

Dependency Structure:

Head noun: کتابی (the book).

acl:relcl: که دیروز خریدم (that I bought yesterday).

Adverbial modifier: دیروز (yesterday).

Main verb: خریدم (bought).

Main verb of the sentence: نخوانده‌ام (haven’t read).

Degree of acl:relcl: 3 (subject, adverbial modifier, verb).

Relative clauses are frequently used in Persian to add detail or clarification, particularly in formal texts or reporting. This leads to a higher degree of acl:relcl because such clauses often include nested or additional descriptive elements.

Deliverable 6

Workload

The workload was distributed based on the team's language fluency and technical strengths. Farhan and Moujan, being fluent in Persian, handled tasks related to Persian text processing and analysis. Their linguistic knowledge was particularly crucial for analyzing dependency parsing, which required a deep understanding of Persian syntax and morphology. Merle, on the other hand, took responsibility for Spanish text processing and implementing the code to analyze tree metrics. While Merle focused on writing and optimizing the scripts, Farhan and Moujan contributed insights about Persian syntax to ensure accuracy in interpreting the results. Collaboration was seamless, with regular updates to ensure consistency in methodology across both languages.

Challenges

One of the key challenges was dealing with the variability in token segmentation, particularly in Persian. Some verbs were overly segmented, while in other cases, sentences were segmented correctly, aligning well with linguistic expectations. This inconsistency required additional preprocessing steps to refine tokenization for a more linguistically coherent analysis. For Spanish, challenges arose in handling contractions and inconsistencies in verb conjugation, which occasionally led to unexpected segmentations. These issues highlighted the need for language-specific adjustments in both tokenization and dependency parsing models.

Surprises

What surprised us most were the unexpected patterns revealed in the dependency tree metrics, such as the average number of degrees, average distance to the root, and most common leave nodes. For instance, in both Persian and Spanish, ccomp (clausal complements) had the highest average number of degrees, as these nodes often hold the detailed structure of subordinate clauses. The ROOT, in contrast, primarily connects high-level components like subjects and objects, leading to a lower degree count. Similarly, in average distance to the root, cc (coordinating conjunctions such as "and" in Spanish or "و" in Persian) had the highest score. This is because conjunctions typically connect clauses or phrases that are far from the root, reflecting their role in joining larger syntactic structures.

In the most common leave nodes, Spanish revealed det (determiners, such as "el" or "la") as a highly frequent category, while this was absent in Persian due to the language's lack of explicit determiners. Additionally, tokenization behavior differed significantly between frequent and infrequent words. Persian, with its complex morphology and use of clitics, led to unexpected segmentations in compound words, while Spanish exhibited inconsistencies in handling verb conjugations and contractions. These observations underscored the importance of adapting parsing and tokenization models to the specific syntactic and morphological properties of the language being analyzed.

References

[1] "Persian Language." *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Persian_language

[2] "All You Need to Know About Persian Language." *Aspirantum*. Retrieved from: https://aspirantum.com/blog/all-you-need-to-know-about-persian-language

[3] "5 Features That Make Persian an Easy Language to Learn." *Tehran Times*. Retrieved from: https://www.tehrantimes.com/news/458617/5-features-that-make-Persian-an-easy-language-to-learn

[4] "The Beauty of Persian: An Introduction for Beginners." *Learn Persian Online*. Retrieved from: https://www.learnpersianonline.com/blog/the-beauty-of-persian-an-introduction-for-beginners/

[5] "Farsi Overview & History." *Study.com*. Retrieved from: https://study.com/academy/lesson/farsi-overview-history-persian-language.html

[6] "Persian Language." *Britannica*. Retrieved from: https://www.britannica.com/topic/Persian-language

[7] "Farsi Language History." *Renaissance Translations*. Retrieved from: https://renaissance-translations.com/farsi-language-history/

[8] "Persian Language." *Asia Society*. Retrieved from: https://asiasociety.org/persian-language

[9] "Leipzig Corpora Collection - Persian Corpus." Wortschatz Leipzig. Retrieved from: https://wortschatz.uni-leipzig.de/en/download

[10] Goldhahn, D., Eckart, T., & Schmid, H. "Building Large Monolingual Corpora." LREC 2012 Proceedings. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

[11] Leipzig Corpora Collection. "Format and Download Information." Retrieved from: https://wortschatz.uni-leipzig.de/public/documents/Format_Download_File-eng.pdf

[12] "UNESCO World Atlas of Languages." UNESCO. Retrieved from: https://en.wal.unesco.org

[13] "World Atlas of Language Structures." WALS. Retrieved from: https://wals.info/feature

[14] "Constitución Española de 1978." *Boletín Oficial del Estado (BOE)*. Retrieved from: https://www.boe.es/buscar/pdf/1978/BOE-A-1978-40001-consolidado.pdf

[15] "Los Problemas de la Lengua Española." *March.es*. Retrieved from: https://www.march.es/es/madrid/problemas-lengua-espanola

[16] "World Factbook: Languages." *CIA*. Retrieved from: https://www.cia.gov/the-world-factbook/countries/world/#people-and-society

[17] "Spanish-Speaking Countries." *Berges Institute*. Retrieved from: https://www.bergesinstitutespanish.com/spanish-speaking-countries

[18] "Standard Spanish." *UNESCO*. Retrieved from: https://en.wal.unesco.org/languages/standard-spanish

[19] "Ethnologue: Spanish." *Ethnologue*. Retrieved from: https://www.ethnologue.com/language/spa/

[20] Green, J. "A Survey of Linguistic Features in Spanish Dialects." *World Atlas of Language Structures (WALS)*. Retrieved from: https://wals.info/refdb/record/Green-1988

[21] Harris, M. "Spanish Phonology and Dialects." *WALS*. Retrieved from: https://wals.info/refdb/record/Harris-1969a

[22] Navarro, T. "Conjugations in Spanish Grammar." *WALS*. Retrieved from: https://wals.info/refdb/record/Navarro-1961
[23] "How is My Spanish? Conjugations." Retrieved from: https://howismyspanish.com/conjugations-in-spanish/
[24] "Exclusión de 'Ch' y 'Ll' del Abecedario." *Real Academia Española (RAE)*. Retrieved from: https://www.rae.es/espanol-al-dia/exclusion-de-ch-y-ll-del-abecedario
[25] "RAE Official Website." *Real Academia Española*. Retrieved from: www.rae.es