


Persian FLAIR: grammatically intelligent web search for language learning

Evan Bartholomeusz^a and Robert Reynolds^b

^aCollege of Life Sciences, Brigham Young University, , ejbart@byu.edu and ^bCollege of Humanities, Office of Digital Humanities, Brigham Young University, , robert_reynolds@byu.edu

How to cite: Bartholomeusz, E.; Reynolds, R. (2023). Persian FLAIR: grammatically intelligent web search for language learning. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik. <https://doi.org/10.4995/EuroCALL2023.2023.16990>

Abstract

We describe our work on the development of a new Persian module for FLAIR, a Java-based grammatically intelligent web search engine (Chinkina et al., 2016; Chinkina & Meurers, 2016). This website allows Persian teachers and Persian language learners to search for texts about any desired topic, prioritizing documents that are rich in selected grammatical constructions, and filtering them based on text difficulty. Currently, the Persian module includes 56 unique grammatical constructions. Our Persian FLAIR module allows for a more self-guided approach to language learning and provides a level of flexibility to both language learners and teachers. Students are able to read accessible texts that interest them, while teachers are able to find real world examples of grammatical concepts used together in the same text. FLAIR can be used to create custom learning materials. Teachers can use FLAIR to search for texts containing particular grammar concepts being focused on in the classroom, which can then be used as relevant topics of discussion or reading. We developed FLAIR in a response to the limited availability of resources for learning and teaching Persian, particularly when it comes to its colloquial forms. With FLAIR, we hope to provide a valuable tool for students and teachers alike.

Keywords: Persian, web search, authentic text, natural language processing.

1. Introduction

Persian, also known by its endonyms of Farsi, Dari, and Tajik, is a less commonly taught language with over 130 million L1 and L2 speakers worldwide (Windfuhr, 2009). It is the official language of Iran, Afghanistan, and Tajikistan, and is also spoken in Uzbekistan, Turkmenistan, and Iraq. Persian continues to hold significance and relevance, not only through its political influence in the Middle East, but also because of the steadily growing Iranian and Afghan diasporas. Despite Persian's worldwide prevalence and importance, it remains a low-resource language. Many universities lack Persian language courses. Most available textbooks focus on formal, prescriptivist approaches and tend to ignore practical applications of colloquial Persian. Current CALL resources for Persian consist mainly of vocabulary apps, with limited resources for pronunciation, grammar, and relevant cultural topics.

Form-focused Linguistically Aware Information Retrieval (FLAIR) is an architecture created by Chinkina, Kannan, and Meurers (2016) to provide learners of English and German with easily accessible comprehensible input in line with the learner's interests. FLAIR uses Microsoft Bing API to fetch results for the user's desired

query, after which the web pages are parsed using Stanford CoreNLP's library (Manning et al., 2014). Research suggests that L2 acquisition occurs best when the learner is made consciously aware of language structure during L2 exposure (Schmidt, 1990). FLAIR allows teachers to find appropriate texts for instruction based on grammatical forms being discussed in the classroom. It also allows learners to independently explore the relationship between form and function in the language without the need for a traditional classroom setting. By developing a similar tool for Persian, we hope to improve the language learning experience for both teachers and learners and to make Persian more accessible to those who are interested in learning.

In this paper, we describe the purpose of FLAIR and specific changes made to adapt its architecture to Persian. Our module is free¹ and open-source².

2. Architecture

FLAIR provides personalizable, interactive input for language learners, which has been shown to facilitate self-correction in L2 production (Gass & Varonis, 1994). Ranking by text difficulty allows language learners to focus on texts best adapted to their current abilities and proficiency levels in the language. Research shows that language learners who are exposed to input at the appropriate difficulty level progress faster than those whose language input is far above or far below their actual language capabilities (Krashen, 1977; Swain, 1985).

FLAIR relies on a combination of machine learning and rule-based categorization to provide users with accurate, relevant search results. By searching for relevant grammatical forms (verb tenses, prepositions, pronoun types, etc.) including those not found in English, users can easily locate texts that showcase desired grammar concepts. This provides them with exposure to natural language constructions used in everyday contexts. FLAIR's architecture then categorizes the texts based on difficulty, allowing users to find level-appropriate learning material. FLAIR is designed to be highly localizable and adaptable to fit any language, and has already proven effective in other languages, including English, German (Chinkina et al., 2016), Russian (Reynolds, 2022), and Arabic (Hveem, 2019). A user selects a search language, enters a query, selects the desired number of results, then the search continues until the desired number of sites with a high enough text content have been found. FLAIR then converts the websites into documents that can be parsed by the selected natural language processing pipeline, which annotates every occurrence of target grammatical constructions.

Once the document is parsed, the user can adjust settings to rank the documents based on the in-text frequency of selected grammatical features, as well as text difficulty based on the CEFR scale. Currently, our analysis relies on a modified Fleisch-Kincaid coefficient, known as Fleisch-Dayani (Dayani, 1990), to determine difficulty level. Users can select multiple grammatical features simultaneously, placing different weights on the ranking of different features. Ranking the texts by grammar concept allows the user to focus on both form and function in the language, observing the interplay between grammar and context in authentic and natural settings. For example, a Persian language teacher covering the Iranian parliament in class, and wanting to teach students about subjunctive verbs, can search for “مجلس ایران” (Iranian parliament), and prioritize texts containing frequent subjunctive verb usages. The teacher can then assign these texts to the students for take-home reading, asking them to highlight or pay attention to each instance of subjunctive verbs and then go online to check Persian FLAIR's highlighted list of all subjunctive verbs to check their work.

3. FLAIR for Persian

Compared to other languages from the Middle East, Persian has a relatively simple grammar. There is no grammatical gender, very few cases, and verb conjugation in all tenses is almost entirely regular. These features lend themselves well to the rule-based aspects of FLAIR's architecture. We created a list of useful grammar

¹ An instance of the server can be accessed at <https://icall.byu.edu/flair-2.0/>

² The source code can be accessed at <https://github.com/reynoldsnlp/flair>. A local instance can be easily deployed using `docker-compose`.

features unique to Persian, adding each feature to a list of grammatical constructions for all supported languages. Each grammatical feature is added to a menu that allows the user to rank and sort each text based on the listed features.

Adding back-end support and text processing for Persian required implementing a new NLP pipeline. Previously supported languages (English, German, Russian, and Arabic) rely on Stanford CoreNLP, which does not currently have models for the Persian language. We implemented Stanza (Qi et al., 2020) as a separate API³, which includes a Persian model based on the Persian Universal Dependency Treebank (Seraji) model for Farsi tagging (Seraji, 2015; Seraji et al., 2016). This required designing a new text processing pipeline inside FLAIR to access the external Stanza API for grammatical parsing. Once the document is processed and tokenized, we are able to extract lemmas, part of speech, grammatical feature tags, and syntactic dependency relations. We then search the grammatical features of each word, attaching labels for identified grammatical constructions. For example, a token categorized as a verb then is searched for grammatical feature tags relating to form, tense, person, negation, and plurality. On the front end of the website, the user then chooses to prioritize documents containing auxiliary verbs and negated verbs, with a higher ranking weight on documents containing negated verbs. The program then sorts the documents and move those containing a higher concentration of the desired feature to the top of the list. In the example below, the query is ranked by texts containing auxiliary verbs and negated verbs. Each result represents a Persian webpage from a Bing search for the bolded term. Based on FLAIR’s user-selected weighted rankings, this query ranks results on quantity of negated verbs. The selected article (the Persian Wikipedia page for the Quran) therefore has the highest quantity of negated verbs, despite other articles potentially containing more auxiliary verbs. (Fig. 1).



Figure 1. A FLAIR query for simple question words and phrases ('Who? What? Where?', bolded text at top of image).

The grammatical constructions included in the Persian module can be broken down into sentence types and parts of speech. Our evaluation of sentence types allows the user to sort by specific question words (who, what, where,

³ Using <https://github.com/lingmod-tue/stanza-api>

when, why, how). We also use the presence of coordinating and subordinating conjunctions to label subordinate clauses and to label sentences as either compound or complex. The presence of relative pronouns is used to label relative clauses.

When sorting by parts of speech, the user is able to select verbs based on verb form (participle adjective, present participle, auxiliary, finite, infinitive, and negated verbs), person, (first, second, and third person, sorted by plural and singular verbs), and tense (simple past, past progressive, simple present, and simple future). Verbs are also sortable by mood (subjunctive and imperative). Quantifiers are sortable by the four most common words (any, some, none, many), as well as by all quantifier words. Adjectives are sortable by degree (positive, comparative, superlative). Adverbs can be categorized as either temporal, locational, negative, or other. Users can rank text by all pronouns, or specify demonstrative, indefinite, interrogative, negative, personal, reciprocal, reflexive, or relative pronouns. Numbers are identified and tagged as either cardinal or ordinal. Nouns are categorized as either plural or singular. Texts can also be ranked by total concentration of all prepositions.

4. Discussion and Conclusions

Persian's literary form has remained largely unchanged for hundreds of years (Jeremias, 2004). Most grammar books and available language learning resources focus on this formal, historical variety, often used in academic reports and traditional poetry. However, the Persian spoken colloquially by roughly 100 million individuals worldwide often bears little outside resemblance to the Persian of textbooks and pedagogy. Colloquial Persian, the language of everyday conversation, online discourse and political discussion, is constantly changing, adapting, and evolving. Mastery of this everyday Persian is absolutely critical for any student who wants to master the Persian language. Yet, when studying Persian, language learners frequently struggle to adapt to colloquial speech and text. The pronunciation and spelling of common verb endings, abbreviations of certain words, and contraction of grammatical markers unique to Persian all become problematic when language learners transition from literary Persian to colloquial texts.

FLAIR's unique text processing capabilities and search functions provide language learners with access to written Persian texts, both colloquial and formal, on any number of desired topics. FLAIR then presents each of these texts through a grammatical lens, showcasing the grammar of the Persian language in a more interactive and engaging way. Beginners can use Persian FLAIR to view simple grammatical constructions and features, first observed in the classroom, in texts that would have otherwise been too difficult for learners at their level. Learners at intermediate and advanced levels can search vast repositories of online information and find level-appropriate texts, sorting and filtering the texts based on grammatical features they are working to master. FLAIR allows users to turn texts of any length, topic, intended audience, or level of formality into a personalized, fully customizable grammar textbook.

Over time, we hope to augment our work on Persian FLAIR to include more grammatical relations, using the dependency tags in the Seraji tree bank to allow the program to examine complicated syntactic relationships in depth. Users can see how each individual word in the sentence relates to the others, highlighting in-context usage of prepositions, grammatical markers, and possessive indicators unique to Persian and how they interact with other aspects of the language. In the future, we also hope to utilize FLAIR's innate localizability to provide native Persian speakers with access to FLAIR's functionality in the currently supported languages of English, German, Russian, and Arabic. Additionally, in future updates, we hope to implement a crowdsourced, machine learning based Persian readability assessment designed by Mohammadi & Khasteh (2020) to more accurately assess text difficulty and give users better rankings.

In conclusion, we have implemented a new Persian module in FLAIR, which allows teachers and learners to filter web search results for 56 grammatical constructions based on lexemes, parts of speech, and other morphosyntactic features. This module will be a useful addition to the Persian teacher's classroom tools, expanding access to this language and its rich history. Persian FLAIR will allow teachers to create more engaging and effective lessons by providing them with a way to target specific grammatical concepts. It will also

help learners to improve their understanding of Persian grammar by giving them access to a wide range of examples. We believe that this will be a valuable resource for both teachers and learners of Persian. We hope that it will help to promote the study of this beautiful and important language, expanding access to the language itself and to its rich history.

Acknowledgements

We would like to acknowledge funding from the Brigham Young University Department of Humanities, Office of Digital Humanities.

References

- Chinkina, M., Kannan, M., & Meurers, D. (2016, August). Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, 7–12.
- Chinkina, M., & Meurers, D. (2016, June). Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 188–198.
- Dayani, M. (1990). A criteria for assessing the Persian texts' readability. *Journal of Social Science and Humanities*, 5(2): 35–48.
- Gass, S., & Varonis, E. (1994). Input, interaction, and second language production. In *Studies in second language acquisition*, 16(03):283–302.
- Hveem, J. (2019). RAFT: Readable Arabic Finding Tool. Unpublished masters project manuscript, Department of Instructional Psychology and Technology, Brigham Young University, Provo, Utah. Retrieved from https://scholarsarchive.byu.edu/ipt_projects/22
- Jeremias, E. M. (2004). "Iran, iii. (f). New Persian". *Encyclopaedia of Islam*. Vol. 12 (New Edition, Supplement ed.). p. 432.
- Krashen, S. (1977). Some issues relating to the monitor model. *On Tesol*, 77 (144–158).
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55-60.
- Mohammadi, H., & Khasteh, S. H. (2020, August). A machine learning approach to Persian text readability assessment using a crowdsourced dataset. In *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, 1–7.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Reynolds, R. (2022). *FLAIR (Russian and Arabic)*. Office of Digital Humanities. <https://odh.byu.edu/projects/flair-russian-and-arabic/>
- Schmidt, Richard W. (1990). The role of consciousness in second language learning. *Applied linguistics*, 11(2):129–158.
- Seraji M. (2015). Morphosyntactic Corpora and Tools for Persian. Doctoral dissertation. *Studia Linguistica Upsaliensia*, 16.

Seraji M., Ginter, F., & Nivre, J. (2016). Universal Dependencies for Persian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2361–2365.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In *Input in second language acquisition*, 15:165–179.

Windfuhr, G. (2009) *The Iranian Languages*, Routledge 2009, p. 418.