Vladimír Benko, Zuzana Kříhová, Boris Lehečka, and Darina Vystrčilová

PERSIAN TO CZECH DICTIONARY
A Traditional Dictionary in the Era of AI?

Abstract We would like to introduce the results of the ELDI project (Electronic Lexical Database of Indo-Iranian Languages, Pilot module: Persian), launched in August 2020. One of the aims of the project was promoting the use of technologies in teaching languages. A website and a mobile application with the Persian–Czech dictionary were developed as the main planned results of the project. A new web-crawled Persian corpus was also created through cooperation with Comenius University in Bratislava and it was primarily used for the study and validation of lexical data, but it is now also directly linked to both applications and is open for use by students, teachers, and researchers in language studies and teaching. The participation of teachers from Charles University in Prague in the project will help to transfer the project outputs into practical teaching. This article also presents the results of a comparison of a dictionary and AI.

Keywords Persian; Czech; lexical database; web-crawled corpus; LLMs

1. Introduction

Persian is a language from the Indo-European language family that is spoken by almost 130 million people, mostly in Iran, Afghanistan, and Tajikistan, representing the three main language varieties of Persian, referred to as Farsi, Dari, and Tajik, respectively. While Farsi and Dari use a modified Arabic script, Tajik is written in modified Cyrillic. Although Czech is an Indo-European language as well, combining it with Persian in the same text pres;ents an additional (typographic) challenge, as these languages differ in the direction of writing.

Our paper introduces some lexicographical and technical aspects of the ELDI Project, the aim of which was to launch a Persian–Czech dictionary website and mobile application (Vystrčilová et al., 2024a,b) and to promote new methods and technologies in teaching non-European languages. In addition, we present one (small-scale) experiment aimed at comparing experiences with traditional and AI-driven Persian to Czech translation.

2. The Lexical Data and the Methods Used to Process Them

The main data source for the Persian module of ELDI was a Persian lexical database that was originally compiled between 1996 and 2020 and served as the basis for three dictionaries that have already been published, two of them in paper form and one on-line (Vystrčilová, 2002; 2014; 2017). The lexical data originally collected on index cards

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

since 1996 were transferred into a text file and then into an Excel spreadsheet, where they were gradually organised into the structure that is described in Tables 1 and 2. After the start of the ELDI project in 2020, the data, divided into several packages, were transferred to *FLEx: FieldWorks Language Explorer* (SIL International, 2024.), where they were further processed, verified, and extended by a larger research team and shared using other tools developed by SIL International (2018): *LanguageDepot* and *FLExBridge*.

At the end of the ELDI project in December 2023, a total of 57,317 dictionary entries were available in the FLEx database file shared by the project team, while 38,312 entries were made accessible in the website and the mobile application. Most of the entries were taken from previous work (52,199 processed entries plus a wordlist of the 24,661 most frequented lemmas) and during the project were subjected to further editing and verification.

The target number of entries is 50,000. The state of verification was the only criterion used for publication, while the frequency of occurrence of the Persian lemma in the mixed language corpora was used just to rank the search results in the ELDI website.

3. Sources of Lexical Evidence

Language corpora have been widely used as the main method for collecting and verifying lexical data since the late 1990s: the first 'home-made' Persian corpus, compiled from the texts of the *Hamshahri* newspaper – probably one of the first Persian open access media published in readable text form and giving readers access to the Persian script – was later replaced by *UPEC* (Uppsala Persian Corpus, Seraji et al., 2012) and the Persian section of *Open Subtitles* (Tiedemann, 2016), using the *WordSmith Tools* corpus manager (Lee et al., 2018; Scott, 2024). In preparation for the ELDI project in 2020, we began to use the *TalkBank Persian* corpus[1] (Rasooli et al., 2013), which could be accessed from the *Sketch Engine* website (Kilgarriff et al., 2014) until April 2022, when its free use was discontinued following the conclusion of the *ELEXIS* Project. The need for a new Persian corpus, therefore, became apparent, and the most straightforward strategy appeared to be to create a new web-crawled corpus.

4. The Persian Web Corpus

During the last decade, the technology for creating corpora from web-crawled texts has been standardised. We were able to use FLOSS tools for most of the components in the processing pipeline (Baroni & Berdnardini, 2004; Pomikálek, 2011; Suchomel & Pomikálek, 2012; Michelfeit et al., 2014), and there were even tools available for the morpho-syntactic annotation of Persian texts, such as *TreeTagger* (Schmid, 1994), *UDPipe* (McDonald et al., 2013; Straka et al., 2016), and *CSTlemma* (Jongejan & Dalianis, 2009), all of which were trained on the *Seraji UD Treebank* (Seraji et al., 2016; Rasooli et al., 2022). As each of the aforementioned tools has its own specific drawbacks, we decided to combine their outputs in the hope of improving the overall

[1] See https://www.sketchengine.eu/talkbank-persian-corpus/; retrieved 21 July 2024.

quality of the annotations. The Persian web space turned out to be quite rich and after just a couple of days of crawling in April 2022 we had managed to create a corpus with over three billion tokens (Benko, 2022), which were processed using the standard pipeline created in the *Aranea* corpora Project.[2]

This corpus is now used by lexicographers as a primary source of lexical evidence for writing dictionary entries, and its 100-million sample is also linked to individual database entries and offers the user the possibility to view selected entries in the context of sentences in the *GDEX* setting (Kosem et al., 2019) or to study them in more detail using the *NoSketch Engine*[3] Corpus Manager (Rychlý, 2007; Kilgarriff et al., 2014).

In the ELDI web application, the relative frequencies in the corpus are represented graphically with green squares, and the entries' frequency values are also used as one of the criteria for ordering the results of the dictionary search.

5. The Structure of the Lexicographic Data

The data structure is summarised in Figure 1 and Tables 1 and 2. Most of the fields come from the source database, but through the collaboration of the project team the database is undergoing further processing. In addition to adding the frequency of occurrence of the entries (at the lemma level) in the large corpora, especially in the newly created *Araneum Persicum*, more example sentences are included from the same source, and each sense definition is assigned to a higher semantic domain according to Moe (2012). More attention is paid to sorting the Czech equivalents into sense categories and their grammatical characteristics, and the system of notes is unified and accompanied by explanations and, for the most part, converted into metadata. Complex expressions (collocations, complex predicates, composites, idioms, etc.) are analysed and linked to relevant components. Chains considered to be variants of the main lexeme, especially stems of irregular verbs, Arabic plurals, complex predicates with alternative light verbs, or orthographic or colloquial variants, are linked to these lexemes. The participation of a native Persian speaker in the project editorial team has contributed significantly to extending the database with idioms and colloquialisms and to the verification of the Czech sense equivalents. Each sense definition is marked with a verification status, which is used to select data for the publicly accessible applications or as a guide for further processing.

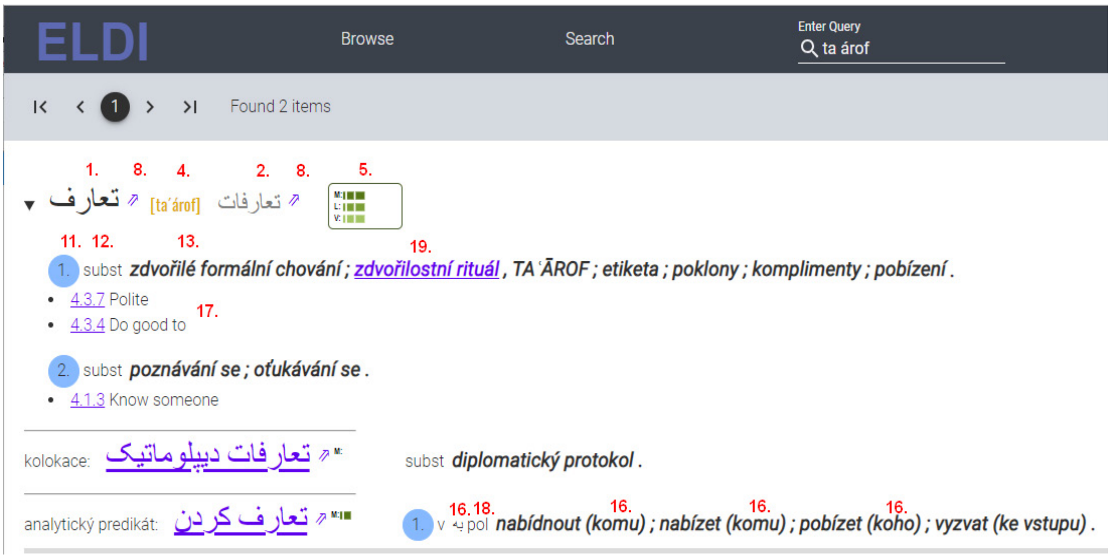
Table 1: Structure of the dictionary entry. A) The headword

No	Field	Comments
1.	Main Headword	The basic form of the entry written in Persian script. This is the second priority field when sorting according to the Persian alphabet.
2.	Variants	Items shown as variants include the Present Stem of Irregular Verbs, Arabic Plurals, Complex Predicates with Alternative Light Verbs, Spelling and Colloquial Variants. The Main Headword and the Variants are represented as interconnected independent entries.

[2] See http://aranea.juls.savba.sk/guest, http://unesco.uniba.sk/guest; retrieved 21 July 2024.

[3] See https://nlp.fi.muni.cz/trac/noske; retrieved 21 July 2024.

3.	Variant Type	Converted to FLEx metadata.
4.	Pronunciation	Transcription of the Headword using the Czech alphabet. It is possible to use the pronunciation to search in the web app.
5.	Frequency	The letters M and T represent the relative frequencies in mixed corpora that meet current standards in terms of size, while the letters L and V represent data from two smaller corpora: UPEC (containing mostly literally language), and the Open Subtitles (spoken language). Six squares represent lexemes with a relative frequency > 1,000 ipm, 5 squares > 100 ipm, etc., and up to one square > 0.01 ipm.
6.	Components	Complex headwords can be divided into components, and seeing the connection between the components and complex forms helps the user to understand the meaning in the same way that the example sentences do.
7.	Complex Forms	Complex Predicates, Collocates, Compounds, Arabic Compounds, Derivatives, Idioms, Sayings.
8.	Links to Concordance	Link to GDEX sentences in a corpus.



▾ تعارف [ta´árof] تعارفات

1. subst *zdvořilé formální chování ; zdvořilostní rituál , TA ʿĀROF ; etiketa ; poklony ; komplimenty ; pobízení .*

- 4.3.7 Polite
- 4.3.4 Do good to

2. subst *poznávání se ; oťukávání se .*

- 4.1.3 Know someone

kolokace: تعارفات دیپلوماتیک subst *diplomatický protokol .*

analytický predikát: تعارف کردن 1. v به pol *nabídnout (komu) ; nabízet (komu) ; pobízet (koho) ; vyzvat (ke vstupu) .*

Fig. 1: Structure of the entry in the ELDI web application[4]

Table 2: Structure of the dictionary entry. B) Sense definition and Czech equivalents

No	Field	Comments
11.	Sense Number	
12.	Grammatical Information	Because of the large number of complex forms in Persian we consider the entry’s semantic function rather than the part of speech. This field is crucial for identifying new senses when importing data into the FLEx.
13.	Czech Equivalents (Definition in FLEx)	The identified Czech Equivalents are organised in the Sense Definition Field in FLEx. They may also contain context notes or indicate prepositional ties. Most of the other types of notes (e.g., register notes) were replaced by metadata taken from the FLEx lists.
14.	Example Sentences	Example sentences compiled mainly from language corpora.

[4] Reverse translation of the Czech senses: Sense 1: polite formal behaviour; courtesy ritual, TAʿĀROF; etiquette; paying respect; compliments; polite encouraging /inviting, Sense 2: getting to know each other. Complex expressions: تعارفات دیپلوماتیک diplomatic protocol, تعارف کردن Sense 1: to offer, to invite (to enter).

15.	Example Translations	Czech translations of the example sentences.
16.	Prepositional Ties (Gram. Note in FLEx)	Prepositional ties and other grammatical and contextual notes. They appear at the beginning of the Definition line.
17.	Semantic Domains	Semantic Domain Thesaurus (Moe, 2012) created by Ron Moe incorporated into the FLEx is used.
18.	Register Notes (Usage in FLEx)	Note on usage in a specific register chosen from the FLEx list Usage.
19.	Hyperlinks to External Sources	Hyperlinks to external sources such as encyclopaedias, e.g., Wikipedia: The free encyclopedia (2024) or Encyclopædia Iranica (Yarshater et al. 1996–).
20.	Cross-references	References to related words or concepts within the dictionary, presenting links between synonyms, etc. (See Fig. 3)

There are other hidden fields in the lexical database that are used, for example, to facilitate the work of the editorial team and to select data for public presentation (Status), to set the alphabetical order or to prioritise search results during the search process (Lexeme Form, Reversals, and Allomorphs), to record sources of study and to verify the particular entry (Bibliography), and to record the history of data processing. The English equivalents were originally used only as a supplementary method of verifying the senses (comparing the English translation from Persian and from Czech), but it is now possible to search data using the sense definitions in English and using the English semantic domains in an advanced search.

6. Data Conversion

The validated data are exported from FLEx to the LIFT standard and then transformed into TEI Lex-0 format (Tasovac et al., 2024), which is being developed as a specialised version of the TEI standard for dictionaries. TEI Lex-0 is being developed by the DARIAH Working Group on Lexical Resources (2024).[5] One goal of this standard is to capture different lexicographic practices in a uniform way and to allow for their comparison and for the mutual linking of existing dictionaries.

For sharing with other applications, the FLEx program exports data in its own XML format (Zook, 2015), with one of the options being the LIFT[6] (Lexicon Interchange FormaT; SIL International, 2024b). The conversion from LIFT format to the TEI Lex-0 standard is fully automated, programmed in the XProc 3.0 programming language (XProc 3.0: Specifications, 2024). The MorganaXProc-IIIse application (Berndzen, 2024) is used to run it. The program converts elements of the LIFT format to corresponding elements of the TEI Lex-0 standard, assigns entries and senses with unique identifiers, and converts semantic category definitions (Moe, 2012) stored in a separate file to the <taxonomy> element in the header of the TEI Lex-0 document, etc. It also generates auxiliary data that the web application uses to present the dictionary, such as translations of data from the meta-language or SVG files, to represent the frequency of lemmas in the corpora.

[5] See https://github.com/DARIAH-ERIC/lexicalresources; retrieved 21 July 2024.

[6] See https://github.com/sillsdev/lift-standard; retrieved 21 July 2024.

The web application uses the XML database eXist-db (eXist Solutions, 2024) version 6.2.0. It comprises two separate components: one is designed for storing and indexing lexicographic data, the other provides access to the stored data through a user interface (web application) and a programming interface (REST API). Both components were created using the TEI Publisher 8.0 framework (e-editiones, 2024)[7] and were later upgraded to version 9.0.[8] This framework makes it possible to create the basic skeleton of the web application with common features, such as a full-text search, the option to browse a document by sections, or the ability to customise the appearance of the presented text using an ODD document.[9] This basic skeleton of the web app can be extended using custom functions (in the XQuery programming language) or components (web pages, JavaScript functions, Web Components.[10] TEI Publisher and related tools use open source code that can be modified.

Among the tools and features of the web application, the following are particularly useful: sorting search results according to the frequency of occurrence of a given lemma, facets enabling the further filtering of search results, superior semantic domains, the hypertext linking of all entries with the new Persian corpus, the mutual linking of orthographic or grammatical variants, components with complex expressions, and, for selected entries, hyperlinks to external encyclopaedic resources.

We paid great attention to the storage of lexicographic data and their indexing. The database contains many indexes in order to make the data and metadata efficiently searchable. The Persian–Czech dictionary in TEI Lex-0 format occupies 120 MB in one file. Indexing such a large file proved to take a disproportionately long time in the order of dozens of hours. This time was reduced by splitting the entire dictionary into separate files, with one file containing one entry. The data were uploaded to the server as a zip file, which also reduced the amount of time required to update the data.

The indexes and their evaluation during query processing affect what data are displayed in the first position as search results. Therefore, for example, when searching for a definition, the resulting score, which will affect the ranking of the results, is constructed based on several criteria scores: a) the frequency of the lemma in the new Persian corpus; b) the position of the searched equivalent in the meaning (equivalents at the top have higher relevance); c) the order of the sense in the entry, and d) the uniqueness of the sense (in which band among all the senses the search term is located).

The source code of the applications referred to above (the conversion of lexicographic data into the TEI Lex-0 standard and two eXist-db modules for data storage and indexing, as well as the user and programming interface) are freely available in the GitHub repository (Lehečka, 2024a-c).

[7] See https://github.com/eeditiones/tei-publisher-app/; retrieved 21 July 2024.

[8] Based on our experience working with the Electronic Database of Indo-Iranian Languages, we created a specialised version that offers features specific to dictionary data: TEI Lex-0 Publisher (Lehečka, 2024d).

[9] 'One Document Does It All' (Text Encoding Initiative, 2024).

[10] Web Components for TEI Publisher (2024), Mozilla (2024).

7. Enhancing Persian and Czech Translation: An Experiment Comparing the ELDI Persian–Czech Dictionary and ChatGPT 4

The integration of the ELDI Persian–Czech Dictionary (FaCsDict) into educational frameworks and translation practices marks a significant advancement in language pedagogy and translation precision. This part of the paper delves into the profound impact of this integration, offering a small-scale analysis of translations using FaCsDict versus those generated by OpenAI's ChatGPT 4. The objective is to illuminate the capabilities and limitations of contemporary AI technologies in translating Persian and Czech and to demonstrate how these tools, when combined with human expertise, can refine translation methodologies and enhance educational outcomes.

The ELDI project is designed to address the underrepresentation of languages such as Persian and Czech in major lexical databases (LDBs). Many existing LDBs prioritise dominant languages like English, leading to a diminished capacity to represent culturally specific words and expressions in less widely spoken languages. By converting traditional dictionary data into a comprehensive Persian-based lexical database using FLEx software, ELDI ensures a detailed and accurate representation of linguistic diversity. This transformation facilitates the preservation and effective translation of culturally specific terms, which is consistent with the goals of improving linguistic representation, as discussed, for example, by Giunchiglia et al. (2023).

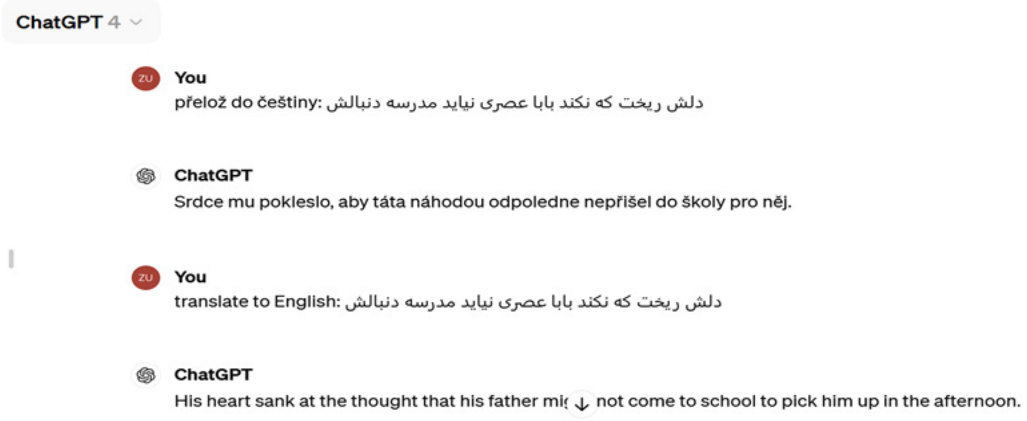
7.1 The Efficiency and Contextual Appropriateness of ChatGPT 4

AI language models, including ChatGPT 4 (OpenAI, 2024), are highly proficient at generating translations rapidly, making them ideal for large-scale translation projects and real-time translation requirements. These models leverage advanced algorithms to consider the text's context, producing translations that are not only fast but also more accurate and contextually appropriate. Recent advances in AI translation models have significantly enhanced multilingual text generation and understanding.

However, despite these advances, AI models often lack the nuanced understanding and cultural sensitivity that human translators possess. This deficiency can lead to potential inaccuracies or misinterpretations, especially in the case of specialised or technical terminology if their training data are not exhaustive (Grassini, 2023). Additionally, AI models often translate through an intermediate language, typically English, before converting the text into the target language. This process can introduce compounded errors, affecting both the intermediate English translation and the final output. For example, in practical translation tasks from Persian to Czech, it has been observed that AI systems may produce misleading or incorrect expressions in the Czech output due to inaccuracies introduced during the English intermediary stage. Such limitations underscore the necessity of integrating human expertise to refine and verify translations, ensuring cultural and contextual accuracy. Below is a sample that illustrates the compounded errors that can occur when AI models translate from Persian to Czech through an intermediate English stage, highlighting the need for human verification.

The specific instance involves the translation of the Persian compound verb دل‌ریختن [del rīkhtan], meaning ‘to get scared’, which was inaccurately rendered. The English translation is acceptable, though not entirely accurate, but the Czech translation is misleading, as it is almost a literal rendition of the English version without considering Czech sentence structure and semantic nuances. The error lies not only in the translation of the idiom (lit. heart drop) into Czech but also in the rendition of the sentence structure into Czech. In Czech, the sentence actually means that the person is afraid that his father might come for him in the afternoon, rather than being afraid that his father won’t come and he will be left alone. The correct translation in Czech is shown in footnotes.

Sample Translation Issue: The following screenshot from ChatGPT 4 demonstrates the compounded errors:



ZU **You**
přelož do češtiny: دلش ریخت که نکند بابا عصری نیاید مدرسه دنبالش

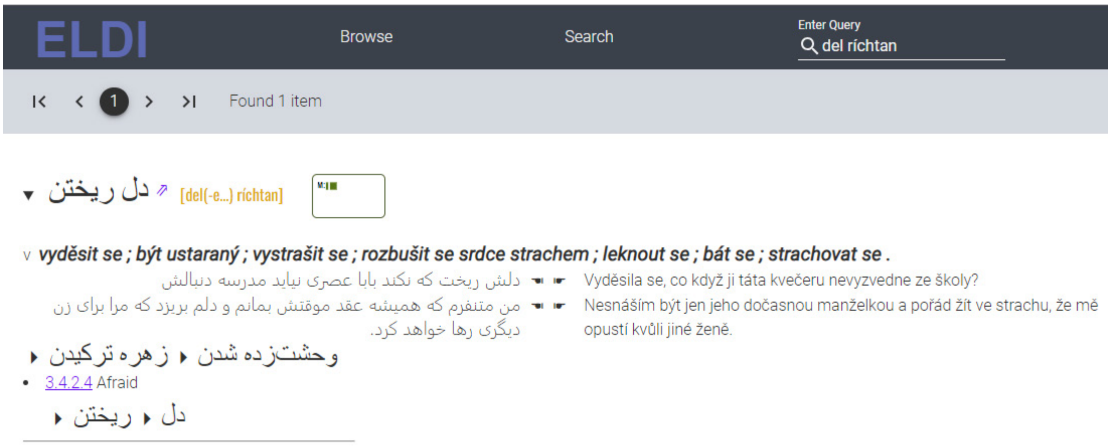
ChatGPT
Srdce mu pokleslo, aby táta náhodou odpoledne nepřišel do školy pro něj.

ZU **You**
translate to English: دلش ریخت که نکند بابا عصری نیاید مدرسه دنبالش

ChatGPT
His heart sank at the thought that his father mi(not come to school to pick him up in the afternoon.

Fig. 2: Screenshot from ChatGPT 4 demonstrating the compounded errors.

To address these issues, the FaCsDict provides a correct translation directly from Persian to Czech, ensuring accurate and contextually appropriate translations. The FaCsDict approach bypasses the intermediate English translation, reducing the potential for compounded errors:



▾ دل ریختن [del(-e...) richtan]

v *vyděsit se ; být ustaraný ; vystrašit se ; rozbušit se srdce strachem ; leknout se ; bát se ; strachovat se .*

دلش ریخت که نکند بابا عصری نیاید مدرسه دنبالش — Vyděsila se, co když ji táta kvečeru nevyzvedne ze školy?

من متنفرم که همیشه عقد موقتش بمانم و دلم بریزد که مرا برای زن دیگری رها خواهد کرد. — Nesnáším být jen jeho dočasnou manželkou a pořád žít ve strachu, že mě opustí kvůli jiné ženě.

‣ وحشتزده شدن ‣ زهره ترکیدن ‣

- 3.4.2.4 Afraid

‣ دل ‣ ریختن ‣

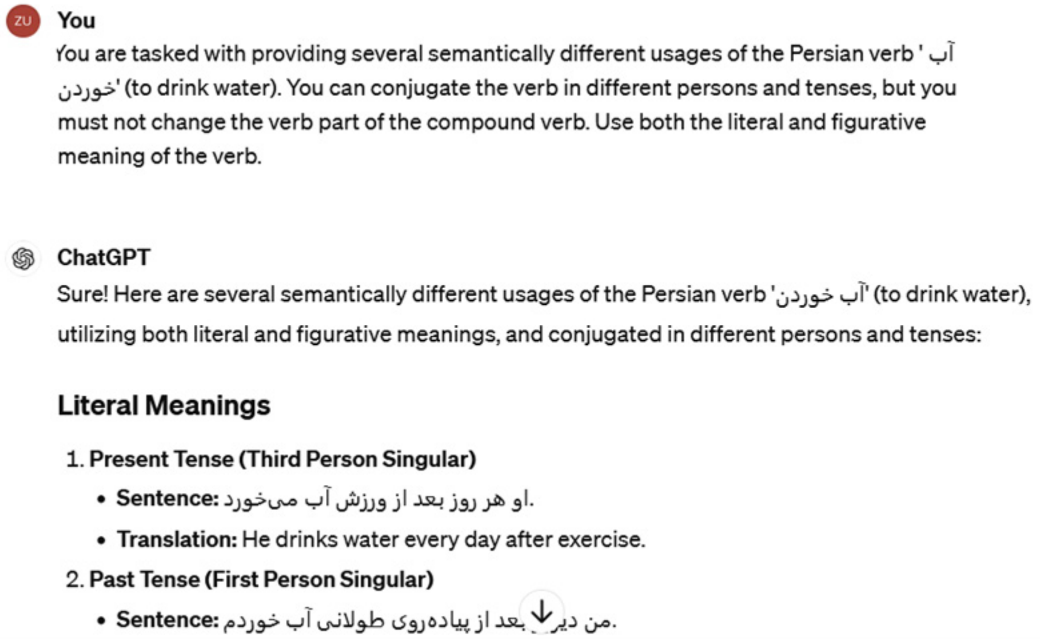
Fig. 3: Translation of the Persian idiom دل ریجتن (lit. ‘heart sink’ by FaCsDict)[11]

[11] Reverse translation of the Czech senses: to get scared; to be worried; to make the heart flutter with fear; to freak out; to fear; to worry. Example: She was scared, what if her dad doesn’t pick her up from school in the evening? I hate being just his temporary wife and always living in fear that he will leave me for another woman.

7.2 Challenges in Translating Polysemy and Idioms

One of the critical challenges in translation is handling polysemy and idioms effectively. AI models like ChatGPT 4, while proficient in generating coherent text, often struggle with these subtleties, leading to translations that may lack the depth and cultural accuracy required for literary texts. This challenge is evident in translations involving expressions with multiple meanings or idiomatic nuances.

For instance, ambiguities or inaccuracies often occur with grammatical constructs related to these issues, as exemplified by the Persian expression آب خوردن [āb khordan], which literally means 'to drink water'. In Persian it also has a range of figurative meanings, which ChatGPT 4 captured in limited and incorrect ways. While the literal translations were accurate, all the figurative examples in Persian were incorrect, and their English translations were likewise inaccurate. In this case, the error made by ChatGPT 4 was more profound and problematic. ChatGPT 4 provided several example sentences by using an existing Persian idiom مثل آب خوردن [mesl-e āb khordan] and (easily) omitting the connection مثل [mesl] (like/as), without which the phrase is meaningless in Persian. Consequently, all the example sentences in Persian are incorrect, even though they seem to provide a comprehensive linguistic sample of the idiom studied. Additionally, while the literal translations were accurate, the figurative meanings that ChatGPT 4 presented in Persian were limited and incorrect. As a result, boththe figurative examples in Persian and their English translations were inaccurate.



You

You are tasked with providing several semantically different usages of the Persian verb ' آب خوردن' (to drink water). You can conjugate the verb in different persons and tenses, but you must not change the verb part of the compound verb. Use both the literal and figurative meaning of the verb.

ChatGPT

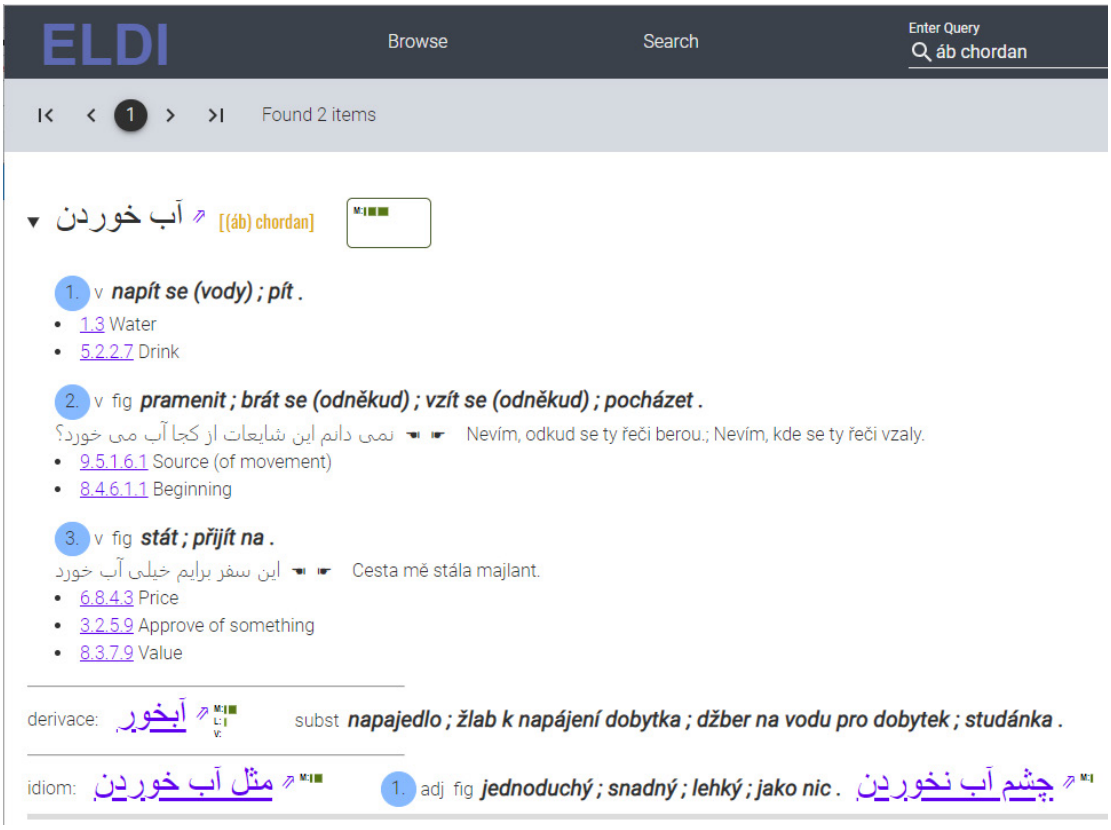
Sure! Here are several semantically different usages of the Persian verb 'آب خوردن' (to drink water), utilizing both literal and figurative meanings, and conjugated in different persons and tenses:

Literal Meanings

1. **Present Tense (Third Person Singular)**
   - **Sentence:** او هر روز بعد از ورزش آب می‌خورد.
   - **Translation:** He drinks water every day after exercise.
2. **Past Tense (First Person Singular)**
   - **Sentence:** من دیر بعد از پیاده‌روی طولانی آب خوردم.

Fig. 4: Screenshot of the ChatGPT 4 translation of the phrase آب خوردن

FaCsDict addresses these complexities by providing contextually rich examples and detailed semantic classifications, ensuring that each word's nuanced meanings are accurately captured and translated. This capability is demonstrated in the following slide, which showcases FaCsDict's effectiveness in accurately translating figurative expressions, highlighting its handling of polysemy and idiomatic phrases compared to AI translation.



▾ آب خوردن ⇗ [(áb) chordan]

1. v *napít se (vody) ; pít .*
- 1.3 Water
- 5.2.2.7 Drink

2. v fig *pramenit ; brát se (odněkud) ; vzít se (odněkud) ; pocházet .*

نمی دانم این شایعات از کجا آب می خورد؟ ☚ ☛ Nevím, odkud se ty řeči berou.; Nevím, kde se ty řeči vzaly.
- 9.5.1.6.1 Source (of movement)
- 8.4.6.1.1 Beginning

3. v fig *stát ; přijít na .*

این سفر برایم خیلی آب خورد ☚ ☛ Cesta mě stála majlant.
- 6.8.4.3 Price
- 3.2.5.9 Approve of something
- 8.3.7.9 Value

derivace: آبخور ⇗ subst *napajedlo ; žlab k napájení dobytka ; džber na vodu pro dobytek ; studánka .*

idiom: مثل آب خوردن ⇗ 1. adj fig *jednoduchý ; snadný ; lehký ; jako nic .* چشم آب نخوردن ⇗

Fig. 5: Translation of the verb/idiom آب خوردن and related complex expressions by FaCsDict[12]

The integration of human expertise is crucial for overcoming the limitations of AI translations. While AI models are adept at handling large volumes of text and providing quick translations, human translators bring the essential skills of cultural sensitivity and contextual understanding. This combination is especially important for languages with less representation in AI training datasets, like Persian and Czech. For instance, idiomatic expressions and nuanced grammatical structures often require a deep understanding of both the source and target cultures, which AI models might not fully possess due to their reliance on statistical patterns rather than lived cultural experiences.

AI models, including ChatGPT 4, are utilised for initial, preliminary translations because of their speed and contextual capabilities. However, the refinement process is supported by FaCsDict, which provides contextually rich examples and detailed semantic classifications. This hybrid approach ensures both efficiency and accuracy, preserving the literary quality of the translated text by accurately capturing the nuanced meanings of polysemous terms and idiomatic expressions. While this hybrid approach ensures both efficiency and accuracy, preserving the literary quality of the translated text by accurately capturing the nuanced meanings of polysemous terms and idiomatic expressions, it should be noted that FaCsDict is still under development and continually being improved. Additionally, using corpus tools is highly effective for verifying the correctness of idiomatic meanings.

[12] Reverse translation of the Czech senses: 1. to drink (water). 2. to originate (from); to come from. Example: I don't know where the rumours are coming from. 3. to cost. The trip cost me a fortune. Complexes: a) watering hole; trough for watering cattle, b) simple; easy; like nothing; a piece of cake.

8. Conclusion

During the ELDI project, a Persian–Czech Database containing some 52,000 processed entries compiled over the period 1996–2020 was imported into a shared FieldWorks Language Explorer file and further extended, modified, and verified by the project team. Over 38,000 dictionary entries chosen by the status of verification were converted into two publicly accessible applications in June 2024: a website using TEI Lex-0 based on TEI Publisher and a mobile app based on Dictionary App Builder. The following features are considered to be its top advantages: content rich in idioms; the direct linking of dictionary entries to GDEX sentences in the corpus Araneum Persicum; solutions to some of the intricacies of Persian orthography; and the inclusion of frequency data to improve the search methods and the ordering of results; the semantic domains assigned to each sense definition, which in combination with the facets (Figure 3 right) and search tools provide users with new ways to explore vocabulary and organize the lexical data.

Significant side results of the ELDI project are a new Persian language corpus called Araneum Persicum and software based on TEI Publisher that is ready for use with other languages. The use of the project results in the teaching of languages is being promoted by the team members at Charles University. Based on a small-scale experiment, the most effective method for educational purposes and translation practices (demonstrated through the example of a literary text in Persian and Czech) involves leveraging the strengths of both AI models and the FaCsDict. The key members of the team are continuing to collaborate on further editorial work as well as consulting on further modifications of the language corpus and the web application.

References

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004* (pp. 1313–1316).

Benko, V. (2022). Aranea Go Middle East: Persicum. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022* (pp. 1–9). Tribun.

Berndzen, A. (2024). *MorganaXProc-IIIse*. Retrieved July 21, 2024, from https://www.xml-project.com/morganaxproc-iiise.html

DARIAH Working Group on Lexical Resources. Retrieved July, 21 2024, from https://www.dariah.eu/activities/working-groups/lexical-resources/

eXist Solutions. (2024). *eXist-db*. Retrieved July, 21, 2024, from https://exist-db.org

FLEx: FieldWorks Language Explorer. Retrieved July, 21, 2024, from https://software.sil.org/fieldworks/

Giunchiglia, F., Bella, G., Chandran Nair, N., Chi, Y., & Xu, H. (2023). Representing Interlingual Meaning in Lexical Databases. *Artificial Intelligence Review*, *56*, 1–17. https://arxiv.org/abs/2301.09169

Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences 13*(7), 692. https://doi.org/10.3390/educsci13070692

Jongejan, B., & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 145–153). Association for Computational Linguistics.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, *32*(2), 119–137.

Lee, S. M., Scott, M., & Shin, D. (2018). *A Practical Approach to Corpus-based Analysis: Introducing WordSmith Tools, Range, and Other Packages.*Lehečka, B. (2024a). *LeDIIR Users.* Retrieved July, 21, 2024, from https://github.com/daliboris/lediir-web-users/

Lehečka, B. (2024b). *LeDIIR Web Application.* Retrieved July, 21, 2024, from https://github.com/daliboris/lediir-web-app/

Lehečka, B. (2024c). *LeDIIR Web Data.* Retrieved July, 21, 2024, from https://github.com/daliboris/lediir-web-data/

Lehečka, B. (2024d). *TEI Lex-0 Publisher.* Version 9.0.1. Retrieved July, 21, 2024, from https://github.com/DARIAH-ERIC/teilex0-publisher

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Z., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, 0., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL* (pp. 92–97). Association for Computational Linguistics.

Michelfeit, J., Pomikálek, J., & Suchomel, V. (2014). Text Tokenisation Using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 71–75). Tribun EU.

Moe, R. (2012). *Semantic Domains.* Retrieved July, 21, 2024, from https://semdom.org

Mozilla. (2024). *Web Components.* Retrieved July, 21, 2024, from https://developer.mozilla.org/en-US/ docs/Web/API/Web_components

OpenAI. (2024). *ChatGPT 4.* Retrieved May, 28, 2024, from https://chat.openai.com

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora.* [Doctoral thesis, Masaryk university, Faculty of informatics].

Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of NAACL-HLT* (pp. 306–314). Association for Computational Linguistics.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *RASLAN* (pp. 65–70). Masaryk University.

Safari, P., Rasooli, M. S., Moloodi, A., & Nourian, A. (2022). The Persian Dependency Treebank Made Universal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7078–7087). European Language Resources Association.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing* (pp. 44–49).

Scott, M. (2008). *WordSmith Tools version 5*. Lexical Analysis Software Ltd. Retrieved July 21, 2024, from http://www.lexically.net/wordsmith/index.html

Scott, M. (2024). *WordSmith Tools version 9* (64 bit version). Lexical Analysis Software.

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

Seraji, M., Megyesi, B., & Nivre, J. (2012). A Basic Language Resource Kit for Persian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2245–2252). European Language Resources Association.

SIL International. (2018). *Technical Notes on FieldWorks Send/Receive*. Retrieved July, 21, 2024, from https://software.sil.org/fieldworks/wp-content/uploads/sites/38/2018/10/Technical-Notes-on-FieldWorks-Send-Receive.pdf

SIL International (2024a). *About SIL*. Retrieved July, 21, 2024, from https://www.sil.org/about

SIL International. (2024b). *Standards*. Retrieved July, 21, 2024, from https://www.sil.org/language-technology/standards

Straka, M., Hajič J., & Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)* (pp. 4290–4297). European Language Resources Association.

Suchomel, V., & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In A. Kilgarriff, & S. Sharoff. (Eds.), *Proceedings of the seventh Web as Corpus Workshop (WAC7)* (pp. 39–43).

Tiedemann, J. (2016). Finding Alternative Translations in a Large Corpus of Movie Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3518–3522). European Language Resources Association.

Tasovac T., Romary L., Banski P., Bowers J., de Does J., Depuydt K., Erjavec T., Geyken A., Herold A., Hildenbrandt V., Khemakhem M., Lehečka B., Petrović S., Salgado A., & Witt, A. (2024). *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.3. DARIAH Working Group on Lexical Resources. Retrieved July, 21, 2024 from https://bit.ly/tei-lex-0

Text Encoding Initiative (2024). *Getting Started with P5 ODDs*. Retrieved July, 21, 2024, from https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/

Vystrčilová, D. (Ed.) (2002). *Příruční slovník persko-český. Farhang-e hamráh-e fársí-čekí* (Persian-Czech Desk Dictionary). PCHE.

Vystrčilová, D. (Ed.). (2014). *Goftegu.cz. Elektronický tematický česko-perský a persko-český slovník.* Retrieved July, 21, 2024, from https://goftegu.cz

Vystrčilová, D. (Ed.). (2017). *Česko-perský slovník. Farhang-e čekí-fársí.* (Czech-Persian Dictionary). Prague.

Vystrčilová, D., Khademi, M., Kříhová, Z., Nachtmann, B., Taucová, R., & Novák, Ľ. (Eds.). (2024a). *Elektronická lexikální databáze indoíránských jazyků: persko-český modul.* Institute of Sociology of the Czech Academy of Sciences, Charles University in Prague. Retrieved July, 21, 2024, from https://eldi.soc.cas.cz

Vystrčilová, D., Khademi, M., Kříhová, Z., Nachtmann, B., Taucová, R., & Novák, Ľ. (Eds.). (2024b). Persko-český mobilní slovník, 2024-06-20. Institute of Sociology of the Czech Academy of Sciences, Charles University in Prague (forthcoming).

Web Components for TEI Publisher. Version 2.24.1. (2024). Retrieved July, 21, 2024, from https://github.com/eeditiones/tei-publisher-components

Wikipedia: The free encyclopedia. (2004). FL: Wikimedia Foundation, Inc. Retrieved July, 21, 2024, from https://www.wikipedia.org

XProc 3.0: Specifications. (2024). Retrieved July, 21, 2024, from https://xproc.org/specifications.html

Yarshater, E., Daniel, E. L., Ashraf, A., Kasheff, M., & Asthiany, M. (Eds.). (1996–). *Encyclopædia Iranica.* Retrieved July, 21, 2024, from https://www.iranicaonline.org

Zook, K. (2015). *FieldWorks 7 XML model.* Retrieved July, 21, 2024, from https://software.sil.org/fieldworks/wp-content/uploads/sites/38/2016/10/FieldWorks-7-XML-model.pdf

Acknowledgements

The project ELDI (TL03000369) was co-financed with the state support of the Technology Agency of the Czech Republic as part of the Éta Program. The development of the Persian web-crawled corpus was, in part, supported by the Slovak VEGA Agency for Science, Project No. 2/0016/21. The contribution of Boris Lehečka in this article in the proceedings was written on the basis of institutional support from the Ministry of Culture for the long-term conceptual development of a research organisation.

The authors thank the anonymous reviewers for their valuable comments, the members of the research team and especially the semantic supervisor Mona Khademi, and Robin Cassling for proofreading this article, the expert consultant in lexicology Michal Škrabal, and the SIL International support team Ann Bush and Ken Zook.

Contact information

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics and Comenius University Science Park, UNESCO Chair in Plurilingual and Multicultural Communication
vladimir.benko@juls.savba.sk

Zuzana Kříhová
Charles University, Faculty of Arts
Zuzana.Krihova@ff.cuni.cz

Boris Lehečka
Moravian Library in Brno
lehecka@mzk.cz, boris@daliboris.cz

Darina Vystrčilová
Czech Academy of Sciences, Institute of Sociology
darina.vystrcilova@soc.cas.cz

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.