*Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 22(1), 1-20.*

# MACHINE-READABLE MULTILINGUAL DICTIONARIES: A SEMI-AUTOMATIC ISO–TEI MODEL[1]

## Makinece Okunabilir Çok Dilli Sözlükler: Yarı Otomatik Bir ISO–TEI Modeli

Emrah ÖZCAN[2], M. Fatih ERKOÇ[3], Hasan TOKATLI[4]

*Yıldız Teknik Üniversitesi*

ORCID ID: 0000-0001-9340-5485[2], ORCID ID: 0000-0002-8278-2805[3], ORCID ID: 0000-0002-9231-4191[4]

**Abstract:** The digital transformation of lexicography has fundamentally shifted the discipline from the production of static, human-readable artefacts to the creation of dynamic, machine-readable databases. This article examines the theoretical and methodological foundations of developing machine-readable multilingual dictionaries (MRDs), with particular emphasis on semi-automatic derivation via pivot languages. Drawing upon the experimental results of a research project we conducted, we analyse the efficacy of utilising the Text Encoding Initiative (TEI) guidelines and ISO-24613 (Lexical Markup Framework) to model complex lexical data. By synthesising historical critiques of database

[1] This article is derived from the General Research Project (GAP) (SBA-2021-4256) carried out with the approval of YTU Scientific Research Projects Coordination Unit.

[2] Yıldız Technical University, Faculty of Education, Department of English Language Teaching, eozcan@yildiz.edu.tr

[3] Yıldız Technical University, Faculty of Education, Computer and Instructional Technologies Education, mferkoc@yildiz.edu.tr

[4] Yıldız Technical University, Informatics Department., htokatli@yildiz.edu.tr

Received: 3 February 2026 ; Accepted: 23 February 2026

How to cite: Özcan, E., Erkoç, M. F., & Tokatlı, H. (2026). Machine-readable multilingual dictionaries: A semi-automatic ISO–TEI model. *Mersin Üniversitesi Dil ve Edebiyat Dergisi*, *22*(1), 1–20.

models with contemporary standards for lexical interoperability, we argue that, while fully automatic induction of multilingual lexicons remains fraught with semantic ambiguity, a semi-automatic workflow, grounded in rigorous data modelling and human verification, offers a scalable solution to overcome the resource scarcity inherent in many language pairs.

**Key words:** *Computational lexicography, Machine-readable dictionaries (MRD), Text Encoding Initiative (TEI), ISO 24613 (LMF), Multilingual dictionaries, Pivot language, Semi-automatic modelling*

**Öz:** Sözlükbilimin dijital dönüşümü, alanı statik ve insan tarafından okunabilir ürünler üretmekten, dinamik ve makine tarafından okunabilir veritabanları geliştirmeye doğru köklü biçimde dönüştürmüştür. Bu makale, özellikle ara (pivot) diller aracılığıyla yarı otomatik türetime odaklanarak, makine tarafından okunabilir çok dilli sözlüklerin (MRD) geliştirilmesine ilişkin kuramsal ve yöntemsel temelleri incelemektedir. Yürüttüğümüz bir araştırma projesinin deneysel bulgularına dayanarak, karmaşık sözlüksel verilerin modellenmesinde Text Encoding Initiative (TEI) yönergeleri ile ISO 24613 (Lexical Markup Framework) standardının kullanım etkinliğini analiz ediyoruz. Veritabanı modellerine yönelik tarihsel eleştirileri, sözlüksel birlikte çalışabilirliğe ilişkin güncel standartlarla sentezleyerek; çok dilli sözlüklerin tamamen otomatik olarak türetilmesinin hâlen anlamsal belirsizliklerle sorunlu olduğunu, buna karşılık sağlam bir veri modellemesine ve insan doğrulamasına dayanan yarı otomatik bir iş akışının, birçok dil çifti için geçerli olan kaynak kıtlığını aşmada ölçeklenebilir bir çözüm sunduğunu savunuyoruz.

**Anahtar sözcükler:** *Hesaplamalı sözlükbilim, Makinece okunabilir sözlükler (MRD), Metin kodlama girişimi (TEI), ISO 24613 (LMF), Çokdilli sözlükler, Ara (pivot) dil, Yarı otomatik modelleme*

## 1. INTRODUCTION

The integration of computational methods into lexicography has been a central pursuit of applied linguistics since the advent of large-scale computing. In the 1980s and 1990s, the field of computational lexicography emerged with the primary goal of converting typesetting tapes of printed dictionaries into Machine-Readable Dictionaries (MRDs) (Boguraev et al., 1990). The objective was to alleviate the "knowledge acquisition bottleneck" in Natural Language Processing (NLP) by extracting ready-made semantic hierarchies and taxonomic structures from existing dictionary data. However, early optimism was

tempered by the realisation that extracting coherent knowledge bases from the idiosyncratic, often inconsistent format of printed dictionaries was significantly more complex than anticipated (Ide & Veronis, 1993).

While early computational lexicography focused on digitizing print sources, the contemporary landscape has shifted toward 'post-editing lexicography.' In this paradigm, the role of the lexicographer evolves from drafting entries to validating data generated by Large Language Models (LLMs) and neural pipelines (Rundell, 2023). As De Schryver (2023) argues, the integration of generative AI has fundamentally altered the compilation process, moving the field toward a 'new age' where the distinction between human and machine authorship is increasingly blurred, necessitating rigorous verification protocols.

Despite these historical challenges, the demand for robust, multilingual lexical resources has intensified, driven by the requirements of the Semantic Web, machine translation, artificial intelligence, and cross-language information retrieval (Aydın et al., 2014). The challenge is particularly acute for “low-resource” languages or language pairs for which direct bilingual dictionaries do not exist. In such contexts, the manual compilation of a multilingual dictionary is often prohibitively expensive and time-consuming (Nasution et al., 2017). Furthermore, the lack of standardisation in legacy data often leads to information “silos” that cannot easily interact with modern NLP tools (Francopoulo & Huang, 2014).

In the research project we conducted, we addressed these challenges by developing a model for the semi-automatic construction of multilingual dictionary systems. We utilised Turkish as a central pivot language to link European languages (English, French, German, and Italian), leveraging open-source bilingual datasets. This article synthesises our findings with established theories of lexical modelling. We posit that the convergence of computational linguistics and traditional lexicography requires a shift from viewing the dictionary as a linear text to viewing it as a hierarchical, graph-based database capable of supporting advanced applications such as cross-language information retrieval and semantic web integration (Vulić et al., 2012; Gillis-Webber, 2018).

### 1.1. RESEARCH QUESTIONS

Responding to the growing need for reusable and interoperable multilingual lexical resources, this paper examines a semi-automatic

approach to generating cross-lingual lexical links through pivot-based induction, using Turkish as a bridge language across multiple TR-centred bilingual dictionaries. Building on the view of dictionaries as structured data models rather than linear texts, the study focuses on workflow design, representational adequacy, and validity risks in induced resources. The study is guided by the following research questions:

1. How can a Turkish-centred pivot-based induction workflow be designed to construct an interoperable multilingual lexical resource from multiple TR-centred bilingual dictionaries?
2. How can the induced multilingual lexicon be modelled and serialised in a TEI/LMF-aligned manner so that lexical information (form, grammatical features, and sense relations) remains consistent and reusable across resources?
3. What are the key semantic risks inherent in pivot-based induction (especially polysemy and multiword phenomena), and what validation and curation strategies (semi-automatic and human-in-the-loop) are most appropriate to minimise error and ensure linguistic plausibility?

## 2. THEORETICAL FRAMEWORK: THE LEXICON AS A DATA MODEL

To render a dictionary useful for both human consultation and machine processing, one must decouple its logical structure from its visual presentation. Traditional dictionaries rely on typographical conventions (boldface, italics, punctuation) to convey structural information implicitly. Computational modelling requires making this structure explicit, a task that historically exposed significant deficiencies in early database models (Boguraev et al., 1990; Ide & Veronis, 1993).

### 2.1. THE HIERARCHICAL NATURE OF LEXICAL ENTRIES VS. RELATIONAL MODELS

Early attempts to model dictionary data often utilised relational database management systems (RDBMS). However, as noted by Boguraev et al. (1990), there is an “inherent conflict between the hierarchical organisation of a dictionary entry and the expressive power of the relational model”. Dictionary entries are naturally tree-like structures: a headword may have multiple homographs; each homograph may have multiple senses; and each sense may contain

definitions, examples, and translations. Flattening this structure into relational tables often leads to massive data redundancy or the loss of scoping relationships, such as the specific link between a sub-definition and a usage note (Boguraev et al., 1990). Furthermore, relational models often struggle to capture the recursive nature of dictionary definitions, in which a sense may contain subsenses to arbitrary depth (Romary & Wegstein, 2012).

Consequently, the field has converged on hierarchical data models, specifically utilising Extensible Markup Language (XML) as the primary vehicle for data representation. XML allows for the nesting of elements (e.g., senses within senses), which mirrors the cognitive organisation of lexical information and allows for the preservation of the "microstructure" of the entry without information loss (Ide & Veronis, 1995).

### 2.2. STANDARDISATION: TEI AND ISO-24613 (LMF)

For a lexical database to be sustainable and reusable, it must adhere to community standards. Two frameworks govern this domain, and their alignment is crucial for modern lexicography:

#### 2.2.1. THE TEXT ENCODING INITIATIVE (TEI)

The TEI Guidelines provide the de facto standard for the digital representation of texts in the humanities. The dictionary module of the guidelines offers a tag set specifically designed to capture the microstructure of dictionary entries, distinguishing between form (<form>), grammatical data (<gramGrp>), and meaning (<sense>) (Ide & Veronis, 1995; TEI Consortium, 2012). The TEI model is flexible enough to accommodate heterogeneous data sources, allowing the representation of both "print-oriented" and "lexical" views within the same document (Lemnitzer et al., 2013).

#### 2.2.2. ISO-24613 (LEXICAL MARKUP FRAMEWORK – LMF)

While TEI focuses on the serialisation of data (the XML format), LMF provides the abstract meta-model. LMF defines the fundamental classes of a lexicon (Lexical Entry, Form, Sense) and their relationships, ensuring semantic interoperability across different projects (Francopoulo & Huang, 2014).

TEI and LMF are structurally isomorphic with respect to the core organisation of lexical entries. Romary and Wegstein (2012) describe the TEI entry as a "crystal", a coherent structure that serialises the abstract classes defined by LMF. By using TEI as the serialisation format for an LMF-compliant model, researchers can achieve both the flexibility of XML and the rigour of ISO standardisation, ensuring that the resulting resource is compatible with the "LMF Reloaded" initiatives (Romary et al., 2019).

The "LMF Reloaded" initiative, driven by the ISO-TC37/SC4/WG4 subcommittee, constitutes a comprehensive review and restructuring of the 2008 standard. The primary motivation behind this revision was to mitigate the inherent complexity of the original model, which, in its attempt to be exhaustive, often lacked sufficient modularity. Furthermore, the 2008 standard notably lacked coverage for diachronic and etymological data and faced challenges regarding compatibility with the widely utilised Text Encoding Initiative (TEI) guidelines (Romary et al., 2019).

### 2.2.3. STRUCTURAL MODULARIZATION

The most significant theoretical shift in the "Reloaded" framework is the move toward high modularity. Rather than a monolithic standard, the new ISO 24613 is divided into multiple distinct parts, allowing users to adopt only the components relevant to their specific research needs. As outlined by Romary et al. (2019), these parts include:

- **ISO 24613-1 (Core Model):** Defines the foundational classes required for a baseline lexicon.
- **ISO 24613-2 (Machine Readable Dictionaries - MRD):** Provides deeper specifications for lexical descriptions, differentiating forms into sub-classes such as Word Form, Stem, and Word Part.
- **ISO 24613-3 (Diachrony-Etymology):** Addresses the previous lacuna regarding word history (discussed further below).
- **ISO 24613-4 (TEI Serialisation):** Explicitly formalises a serialisation based on a restricted version of the TEI guidelines.
- **ISO 24613-5 (LBX Serialisation):** Provides a serialisation using the Language Base Exchange format.
- **ISO 24613-6 (Syntax and Semantics):** Revises syntactic and semantic components.

• **ISO 24613-7 (Morphology):** A separate module dedicated to morphological patterns.

From a data modelling perspective, "LMF Reloaded" introduces crucial simplifications. One notable improvement is the introduction of the *CrossREF* class. In previous iterations, modelling Multi-Word Expressions (MWEs) required complex structures involving a *List of Components* and *Component* classes. The *CrossREF* mechanism replaces these with a streamlined pointing mechanism that models semantic relations, cross-references, and related entries more efficiently (Romary et al., 2019).

Simultaneously, the model has been enriched to support granular linguistic inquiry. For instance, the initiative differentiates *Orthographic Representation* into *Form Representation* and *Text Representation*, thereby allowing greater precision in encoding written forms at the *Sense* and *Form* levels (Romary et al., 2019).

## 3. METHODOLOGY: THE PIVOT LANGUAGE APPROACH

The core methodological contribution of the project we conducted is the use of a "pivot" or "bridge" language to induce new bilingual pairs from existing resources. This technique is essential when creating dictionaries for language pairs where direct resources are scarce.

### 3.1. LEXICAL RESOURCES AND DATA SOURCES

The multilingual resource developed in this project is based on four open-source, XML-encoded bilingual dictionaries obtained from the FreeDict repository. These include English–Turkish (36,589 headwords), French–Turkish (10,869 headwords), German–Turkish (36,219 headwords), and Italian–Turkish (17,932 headwords). All resources are distributed in a structured XML format compatible with TEI-inspired lexicographic modelling, enabling uniform parsing and integration. Turkish was selected as the pivot language due to its availability across multiple bilingual resources and its capacity to function as a semantic bridge between Indo-European languages. The use of openly licensed, standards-compliant datasets ensures both reproducibility and extensibility of the proposed methodology.

### 3.2. TRANSITIVE INFERENCE AND CONSTRAINTS

The logic of pivot-based induction is transitive. Let L_A be the source language, L_B be the pivot language, and L_C be the target language. If word w_A in L_A translates to w_B in L_B, and word w_C in L_C translates to w_B in L_B, then we can infer a translation link between w_A and w_C via the pivot w_B.

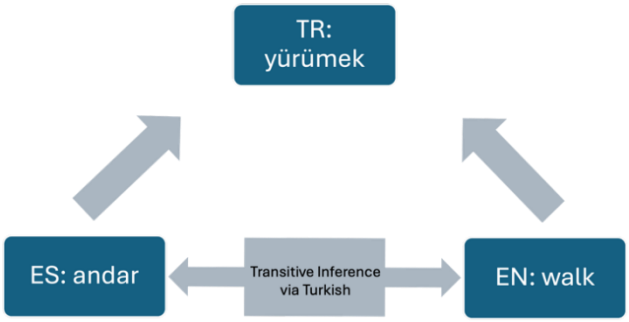


**Figure 1.** Pivot-based transitive inference in multilingual lexical alignment.

In our project, Turkish served as the pivot (L_B). We utilised four open-source bilingual dictionaries (English-Turkish, French-Turkish, German-Turkish, and Italian-Turkish) to generate a unified multilingual matrix. Figure 1 illustrates transitive inference via a pivot language: a source-language item (EN: *walk*) and a target-language item (ES: *andar*) are linked indirectly through their shared translation to a pivot-language equivalent (TR: *yürümek*), enabling the induction of a translation relation between the source and target under pivot-based constraints.

This approach allows for rapid expansion of lexical coverage. As Nasution et al. (2017) note, constraint-based pivot techniques are particularly effective for low-resource languages, provided that suitable filtering mechanisms are used to manage ambiguity. Similar methods have been successfully employed using parallel corpora to generate bilingual lexicons, as shown in the development of English-Macedonian resources (Saveski & Trajkovski, 2011), reinforcing the validity of transitive induction when direct translation resources are not available.

This pivot-based configuration enables a substantial expansion of multilingual lexical coverage from a limited set of bilingual resources. As summarised in Table 1, four Turkish-centred dictionaries (English-Turkish, French-Turkish, German-Turkish, and Italian-Turkish) serve as the basis for inducing six additional bilingual lexicons without requiring any direct language-pair data. The table reports both the size of the original resources and conservative lower-bound estimates of the number of entries inferred for each induced language pair. By constraining induction to shared, normalised Turkish pivot entries, the approach maintains linguistic grounding while achieving scalability, with all inferred links subsequently treated as candidates for layered human validation.

**Table 1.** Direct bilingual resources used in the project and the approximate number of bilingual translation links induced via Turkish as a pivot language under transitive inference constraint[5].

| *Language Pair* | *Resource Type* | *Headwords* | *Role in Induction* |
|---|---|---|---|
| English–Turkish | Direct bilingual | 36,589 | Pivot-linked source |
| French–Turkish | Direct bilingual | 10,869 | Pivot-linked source |
| German–Turkish | Direct bilingual | 36,219 | Pivot-linked source |
| Italian–Turkish | Direct bilingual | 17,932 | Pivot-linked source |
| English–French | Induced | ≈ 5,400 | Pivot-based inference |
| English–German | Induced | ≈ 18,000 | Pivot-based inference |
| English–Italian | Induced | ≈ 9,000 | Pivot-based inference |
| French–German | Induced | ≈ 5,400 | Pivot-based inference |
| French–Italian | Induced | ≈ 5,400 | Pivot-based inference |
| German–Italian | Induced | ≈ 9,000 | Pivot-based inference |

[5] Estimation logic**:** Induced bilingual entries were estimated by intersecting normalised Turkish pivot lemmas across each language pair. Only exact or morphologically normalised matches were counted. To avoid inflation due to polysemy, each pivot lemma was assumed to contribute at most one candidate entry per induced pair, yielding conservative lower-bound estimates.

### 3.3. DATA PROCESSING AND "SEMI-AUTOMATIC" COMPILATION

In our project, we used a semi-automatic workflow. The term "semi-automatic" is important because fully automatic extraction often yields high error rates due to polysemy (as discussed in Section 5). The workflow in our project included:

**Ingestion:** Parsing disparate bilingual source files (often in simple text or CSV formats) into a standardised TEI-XML structure.

**Alignment:** Using SQL-based algorithms to match entries based on the Turkish pivot column. This required normalising the data to ensure that morphological variations in the pivot language did not obstruct matching.

**Verification:** Allowing human experts to verify the suggested links. This "human-in-the-loop" approach mitigates the risks of automated erroneous linking, ensuring that cultural nuances and specific sense distinctions are preserved.

## 4. STRUCTURAL MODELLING: THE TEI DICTIONARY ARCHITECTURE

The success of a machine-readable dictionary depends on the granularity of its markup. In our project, we adopted a rigorous TEI structure that segments the dictionary microstructure into discrete, machine-processable components, aligning with the "crystal" structures proposed by Romary and Wegstein (2012).

### 4.1. THE MICROSTRUCTURE

We modelled the dictionary using a hierarchical XML tree rooted in the <entry> element. Following the TEI P5 guidelines, the entry is divided into three primary components, mirroring the LMF core package:

**Form Information (<form>):** As seen in example (1), the <form> element constitutes the outer layer of the lexical entry and serves as the container for all information related to the signifier, that is, the observable and phonetic realisation of the lexical unit. It encapsulates orthographic representation (<orth>) and pronunciation (<pron>), allowing these dimensions to be encoded independently of grammatical or semantic information. In the database generated by our project, for instance, the French lexical item *rémoulade* is encoded with its standardised phonetic transcription nested within the <form> element. This explicit separation supports interoperability and is particularly

advantageous for downstream applications such as speech synthesis, automatic speech recognition, and pronunciation modelling, where access to phonological data must be direct and unambiguous (Sobkowiak, 1996).

(1)
```
<form type="lemma">
  <orth>rémoulade</orth>
  <pron notation="ipa">ʁemulad</pron>
</form>
```

**Grammatical Information (<gramGrp>):** Positioned as a distinct component within the lexical entry structure shown in example (2) below, the <gramGrp> element aggregates morphosyntactic features associated with the lexical item and grouped within the form block in our implementation. In our implementation, this includes standardised descriptors such as <pos> (part of speech) and <gen> (gender), enabling consistent grammatical annotation across entries. By isolating grammatical metadata from definitional or contextual text, the model ensures computational accessibility, enabling morphological parsers and taggers to retrieve grammatical information without requiring text mining or heuristic inference. This design choice reflects established best practices in computational lexicography and contributes to efficient linguistic processing and reuse across NLP pipelines (Ide & Veronis, 1995).

(2)
```
<form type="lemma">
  <orth>rémoulade</orth>
  <gramGrp>
    <pos>noun</pos>
    <gen>feminine</gen>
  </gramGrp>
</form>
```

**Sense Information (<sense>):** As represented at the semantic level of the entry, the <sense> element functions as the core unit for meaning representation within the lexicon. Crucially, the TEI framework

permits recursive nesting of <sense> elements, a feature that is central to modelling semantic complexity. This structure enables the explicit representation of polysemy, where a single lexical form corresponds to multiple meanings, as well as subsenses that capture finer semantic distinctions. Such hierarchical organisation, shown in example (3), supports both human-readable lexicographic description and machine-readable semantic processing, facilitating tasks such as word sense disambiguation and semantic alignment across resources (Ide et al., 2000; Romary & Wegstein, 2012)

(3)
```
<sense n="1">
  <def>A sauce made from mayonnaise, herbs, and seasonings.</def>
  <sense n="1.1">
    <def>A regional variant using mustard as a base.</def>
  </sense>
</sense>
```

Figure 2 illustrates the complete set of structural components required for a lexical entry in the proposed model, encompassing form, morphological, and sense-related information. Each component group contains the essential descriptive elements needed to represent orthographic, phonological, grammatical, and semantic properties of a lexical item, together with their cardinality constraints. The organisation of these elements is fully aligned with the Text Encoding Initiative (TEI) P5 guidelines, ensuring that the model conforms to established standards for lexical encoding and supports interoperability, consistency, and machine readability across lexicographic resources. This structured separation of form, grammatical information, and sense representation is critical for pivot-based alignment, as it allows transitive inference to operate on normalised lexical units rather than surface strings.

Current standards require that TEI-encoded data be interoperable with the Semantic Web. The integration of TEI Lex-0 with the OntoLex-Lemon model facilitates the publication of legacy dictionaries as Linked Open Data, allowing for collaborative editing on platforms such as Wikibase (Lindemann, 2025; Krek et al., 2025). This interoperability is crucial for non-Indo-European languages as well; for instance, Jarrar and Amayreh (2019) successfully utilized these

standards to map Arabic ontologies, demonstrating the model's cross-linguistic validity.

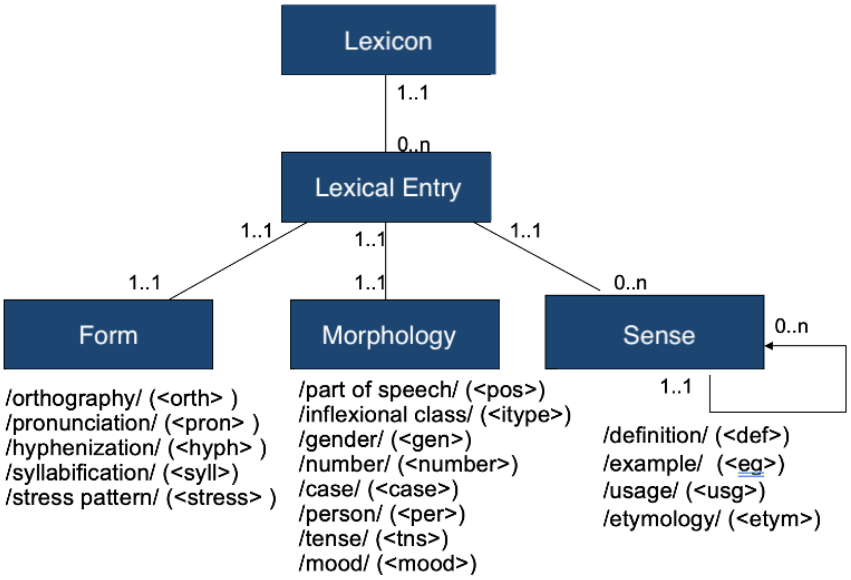


**Figure 2.** Representing lexical entries through form, morphology, and sense components in line with TEI-inspired lexicographic modelling

### 4.2. QUANTITATIVE LEXICAL YIELD

The application of pivot-based induction to the four Turkish-centred dictionaries results in a substantial expansion of multilingual lexical coverage. On the basis of conservative estimates derived from shared, normalised Turkish pivot lemmas, the system induces approximately 50,000-55,000 indirect bilingual entry candidates across the six inferred language pairs (see Table 1). These figures represent lower-bound estimates, as induction is restricted to entries exhibiting identical or morphologically normalised pivot forms. All inferred links are therefore treated as provisional and subsequently subjected to layered validation procedures, ensuring that scalability is balanced with semantic accuracy. This quantitative yield reflects the underlying modular TEI/LMF data architecture, in which the explicit separation of form and sense representations enables reliable transitive alignment across languages.

## 5. CHALLENGES IN MULTILINGUAL EXTRACTION

While the pivot approach is efficient, it introduces specific semantic noise that must be managed. The literature on extracting knowledge bases from MRDs identifies "tangled hierarchies" and polysemy as the primary obstacles. The scale of the induced lexicon further underscores the necessity of a semi-automatic workflow. While the automatic phase enables the generation of tens of thousands of candidate entries, semantic ambiguity, particularly arising from polysemous pivot forms, cannot be reliably resolved without human intervention. Consequently, the project adopts a layered human-in-the-loop strategy, in which automatically generated candidates are first filtered through large-scale, low-cost validation mechanisms and subsequently refined through review.

### 5.1. THE POLYSEMY PROBLEM

The most significant risk in pivot-based induction is the "polysemy trap". If the pivot word in the pivot language is polysemous, it may incorrectly link unrelated words in the source and target languages.

*Scenario:* Turkish word *yüz* can mean "face" or "hundred" depending on the context. If English *face* maps to *yüz*, and German *hundert* maps to *yüz*, a naive algorithm might infer that English *face* translates to German *hundert*.

**Historical Context**: Studies by Veronis and Ide (1991) have shown that single-dictionary extraction can result in error rates of 55-70% due to such ambiguities. They demonstrated that merging data from multiple dictionaries significantly reduces these errors by filtering out inconsistencies.

**Mitigation:** Advanced methods, such as Inverse Consultation (IC), aim to mitigate these errors by performing a reverse check (i.e., translating from C to A) (Nasution et al., 2017). Building on the semi-automatic workflow described in Section 3.3, the algorithmic output serves as a candidate list for human verification and requires careful manual handling.

Beyond single-word polysemy, additional ambiguity arises from multiword units, which introduce non-compositionality and discontinuity that are difficult to capture in standardised lexical representations. A persistent challenge in semi-automatic extraction remains the identification of Multiword Expressions (MWEs), which

often defy compositional translation. A comprehensive recent survey of lexical resources indicates that while detection techniques have improved, the representation of MWEs in standard formats like OntoLex-Lemon requires specialized properties to capture non-compositionality and discontinuity (Chiarcos et al., 2024).

In line with the semi-automatic workflow described above, the bulk of processing is carried out automatically via algorithms that generate candidate structures. These automatically produced outputs are then subjected to a carefully designed human verification and correction phase, in which domain experts review, disambiguate, and validate the results. This division of labour ensures both efficiency and semantic accuracy: automation enables broad coverage and scalability, while targeted manual intervention addresses ambiguity, polysemy, and noise that cannot be reliably resolved by algorithms alone.

## 6. FUTURE DIRECTIONS: LINKED DATA AND THE SEMANTIC WEB

From a forward-looking perspective, the creation of a standardised, TEI/LMF-compliant XML database should be understood not as an endpoint of lexicographic digitisation but as a transition toward the Semantic Web and next-generation AI-driven language technologies. While the pivot-based modelling strategy employed in this project yields substantial quantitative gains, enabling the transformation of a limited set of bilingual dictionaries into a densely connected multilingual lexical matrix, its broader significance lies in the semantic transparency and interoperability of the resulting data structures. By encoding lexical entries according to internationally recognised standards, the resource remains firmly grounded in lexicographic theory while simultaneously becoming accessible to computational systems that require explicit, machine-interpretable representations of form, meaning, and grammatical relations.

Since TEI/LMF models are inherently compatible with Linguistic Linked Data frameworks, the lexicon can be seamlessly integrated into distributed semantic networks and reused across heterogeneous platforms. This compatibility enables lexical data to serve not only as a reference resource for human users but also as structured input for AI pipelines that support tasks such as semantic reasoning, cross-lingual inference, and knowledge representation. In this sense, the multilingual matrix produced through pivot-based induction serves as both a lexicographic artefact and a training-ready semantic resource, capable

of contributing to data-driven fine-tuning and evaluation of language technologies that depend on machines' increasingly sophisticated understanding of textual and lexical information.

### 6.1. THE RESPONSIVE DICTIONARY AND CROWDSOURCING

The ultimate goal of digitising these resources is to create responsive dictionaries. As defined by Gantar (2020), a responsive dictionary actively involves the user community in the lexicographic process, most notably through crowdsourcing mechanisms that help identify and reduce the noise generated by automatic extraction. In our project, this human involvement is conceived as a layered and phased workflow rather than a single manual intervention. At an initial layer, non-expert users contribute through crowdsourced validation tasks—such as flagging errors, confirming basic equivalences, or suggesting corrections—thereby providing large-scale, low-cost feedback. This is followed by a second, expert-level layer in which trained lexicographers review, adjudicate, and refine the data to ensure terminological consistency and semantic accuracy.

Crucially, the project's ultimate objective is to leverage these human-in-the-loop layers not only for correction but also for fine-tuning the system's automatic components. Feedback gathered at each stage is intended to inform iterative model improvement, progressively reducing noise in subsequent extraction and alignment cycles. The XML infrastructure developed in our project is well suited to import into open-source dictionary-writing systems, such as Lexonomy, which support collaborative editing and publishing (Mechura, 2017). Furthermore, tools such as GROBID-Dictionaries increasingly employ machine learning techniques to automatically structure legacy PDF dictionaries into TEI-compliant formats, thereby accelerating the population of these systems and reinforcing a continuous cycle of automation, human validation, and model refinement (Khemakhem et al., 2018).

The frontier of lexicography is currently defined by the application of Large Language Models (LLMs) for data enrichment. Pilot studies at the Dutch Language Institute have utilized LLMs to generate definitions and validate semantic shifts, though results indicate that human expert verification remains essential to avoid hallucination (Tiberius et al., 2024). Furthermore, LLMs are being adapted for historical dialect lexicography, where they assist in the semantic classification of non-standardized lemmas, a task previously deemed too complex for automation (Stöckle et al., 2025).

## 7. CONCLUSION

The research conducted in our project provides substantial validation of a standards-based approach to computational lexicography, demonstrating how formally defined models can support both scalability and semantic control. By employing Turkish as a pivot language, we constructed a multilingual lexical resource from pre-existing bilingual datasets in a manner that is both extensible and computationally efficient. The findings confirm that, although automatic extraction and alignment techniques have advanced considerably since early discussions of *tangled hierarchies* and structural noise in machine-readable dictionaries (Veronis & Ide, 1991), fully unsupervised solutions remain limited by polysemy and semantic ambiguity. Consequently, a *human-in-the-loop* paradigm remains indispensable. In our workflow, the majority of processing is carried out automatically, while human intervention is organised into layered phases—ranging from large-scale, crowdsourced validation to expert-level lexicographic review—thereby allowing manual effort to be applied selectively and strategically.

Quantitatively, the project demonstrates that a small number of well-structured bilingual resources can be leveraged to produce a disproportionately large multilingual lexicon when combined with pivot-based transitive inference. The induction of over fifty thousand candidate entries from four Turkish-centred dictionaries provides empirical evidence that semi-automatic lexicon construction can meaningfully address lexical resource scarcity while remaining compatible with future AI-oriented semantic infrastructures.

The adoption of ISO 24613 (LMF) and TEI guidelines ensures that the resulting data is not merely a self-contained digital artefact but a structurally interoperable component within the broader ecosystem of linguistic linked data (Romary et al., 2019). This standards compliance enables seamless integration with Semantic Web and AI-oriented language technologies, positioning the resource beyond traditional lexicographic use and toward machine understanding of text. By decoupling the underlying data model from the presentation layer and combining semi-automatic pivot-based induction with rigorous data modelling and human verification, the project demonstrates a viable pathway to address the chronic resource gaps affecting many underrepresented languages, while simultaneously laying the

groundwork for future AI-driven lexical and semantic applications. This aligns with the vision of intelligent lexicography (aiLEX), where the resource adapts to the user profile rather than offering a "one-size-fits-all" interface (De Schryver, 2010).

The transition from static text to structured TEI/LMF data transforms the dictionary from a passive reference tool into a dynamic educational tool. By structuring lexicographic data and separating *form*, *sense*, and *usage*, the goal is to achieve an "adaptive" dictionary model that can present different data views depending on the learner's proficiency (e.g., filtering complex morphology for beginners). The explicit encoding of morphological anomalies and collocational patterns allows for the automatic generation of exercises or the identification of "defective" forms that learners need to be warned about (Kovarikova, 2021).

High-quality, structured lexical data is the prerequisite for educational language games. As demonstrated by the CJVT Igre platform, structured dictionaries allow for the automatic generation of semantic puzzles and orthographic challenges, which significantly enhance vocabulary retention and engagement for younger learners (Arhar Holdt & Kosem, 2025; Rabe et al., 2025). Our multilingual dictionary model could help developing new Computer-Assisted Language Learning (CALL) applications as well.

REFERENCES

Arhar Holdt, S., & Kosem, I. (2025). Using large language models to generate distractors for language games. In *Proceedings of the eLex 2025 Conference* (pp. 620–635). Bled, Slovenia.

Aydın, C. R., Erkan, A., Güngör, T., & Takçı, H. (2014). Sözlük kullanarak Türkçe için kavram madenciliği metotları geliştirme: Bir uygulama. In *Proceedings of XVI. Academic Informatics Conference* (pp. 801–810). Mersin, Türkiye.

Boguraev, B., Briscoe, T., Carroll, J., & Copestake, A. (1990). Database models for computational lexicography. In *Proceedings of the 4th International Congress on Lexicography* (pp. 59–78). Malaga, Spain.

Chiarcos, C., Ionov, M., Apostol, E.-S., Gkirtzou, K., Kabashi, B., Khan, A. F., & Truică, C.-O. (2024). *Multiword expressions, collocations and the OntoLex vocabulary*. In Multiword expressions in lexical resources, 187-227. Language Science Press. https://doi.org/10.5281/zenodo.10998641

De Schryver, G.-M. (2010). State-of-the-art software to support intelligent lexicography. In R. Zhu (Ed.), *中華字典研究-第2輯(上下)-2009《康熙字典》暨詞典学国際学術研討会論文集* 2, 584–599. 中国社会科学 = China Sociale Wetenschappen Publishing House.

De Schryver, G.M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355-387.

Francopoulo, G., & Huang, C.-R. (2014). Lexical markup framework: An ISO standard for electronic lexicons and its implications for Asian languages. *Lexicography ASIALEX, 1*, 37-51. https://doi.org/10.1007/s40607-014-0006-z

Gantar, P. (2020). Dictionary of Modern Slovene: From Slovene lexical database to digital dictionary database. *Rasprave*, 46(2), 589-602.

Gillis-Webber, F. (2018). Conversion of the English-Xhosa dictionary for nurses to a linguistic linked data framework. *Information*, *9*(11), 274. https://doi.org/10.3390/info9110274

Ide, N. & Veronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time?, *Knowledge Bases & Knowledge Structures 93*, Tokyo, 257-266.

Ide, N., & Veronis, J. (1995). Encoding dictionaries. In N. Ide and J. Veronis (Eds.), *The Text Encoding Initiative: Background and Context*, *special triple issue of Computers and the Humanities*, *29*(2), 167-180.

Ide, N., Kilgarriff, A., & Romary, L. (2000). A formal model of dictionary structure and content. *Euralex 2000 Proceedings*, 113-126.

Jarrar, M., & Amayreh, H. (2019). An Arabic-Multilingual Database with a Lexicographic Search Engine. In *Lecture notes in computer science*, 234-246. Springer International Publishing. https://doi.org/10.1007/978-3-030-23281-8_19

Khemakhem, M., Herold, A., & Romary, L. (2018). Enhancing usability for automatically structuring digitised dictionaries. *GLOBALEX workshop at LREC 2018, May 2018*, Miyazaki, Japan. https://hal.science/hal-01708137v1

Kovarikova, D. (2021). Sharing data through specialized corpus-based tools: The case of GramatiKat. *Journal of Linguistics/Jazykovedný casopis. 72*, 531-544. https://doi.org/10.2478/jazcas-2021-0049.

Krek, S., Ponikvar, P., Repar, A., Kosem, I., and Lindemann, D. (2025). DMLEX on Wikibase: Legacy dictionaries as collaboratively editable dataset *Proceedings of the eLex 2025 conference*, Bled, Slovenia, 175-189.

Lemnitzer, L., Romary, L., & Witt, A. (2013). Representing human and machine dictionaries in markup languages. In R. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *HSK - Dictionaries. An international encyclopedia of lexicography: Supplementary volume: Recent developments with special focus on computational lexicography* (Vol. 5.4, pp. 1195–1208). Mouton de Gruyter.

Lindemann, D. (2025). Ontolex-Lemon in Wikidata and other Wikibase instances. *Proceedings of the 5th Conference on Language, Data and Knowledge: Workshops*, 287–297. https://doi.org/10.5281/zenodo.15861038

Mechura, M. (2017). Introducing Lexonomy: An open-source dictionary writing and publishing system. *Proceedings of eLex 2017 Conference*, 662-679.

Nasution, A. H., Murakami, Y., & Ishida, T. (2017). Plan optimization for creating bilingual dictionaries of low-resource languages. *Proceedings of IEEE International Conference on Culture and Computing*, 35-41.

Rabe, M., Puttkammer, M. J., & van Huyssteen, G. B. (2025). Compiling a Candidate List of Taboo Constructions for an Under-Resourced Language. *Proceedings of the eLex 2025 conference*, 739-756.

Romary, L., & Wegstein, W. (2012). Consistent modeling of Heterogeneous Lexical Structures. *Journal of the Text Encoding Initiative*, *3*, 1-43.

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M., & Bański, P. (2019). LMF reloaded. *Proceedings of the AsiaLex 2019 Conference*, 533–539.

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? *proceedings of the 16th International Conference of the Asian Association for Lexicography (ASIALEX 2023)*, 9-17.

Saveski, M., & Trajkovski, I. (2011). Development of an English–Macedonian machine readable dictionary by using parallel corpora. In M. Gusev & P. Mitrevski (Eds.), *ICT Innovations 2010. Communications in Computer and Information Science* (Vol. 83, pp. 207–218). Springer. https://doi.org/10.1007/978-3-642-19325-5_20

Sobkowiak, W. (1996). Phonetic Transcription in Machine-readable Dictionaries. *Proceedings of the 7th EURALEX International Congress (EURALEX '96)*, 181-188.

Stöckle, P., Elsner, D., Koppensteiner, W., & Korecky-Kröll, K. (2025). LLM-assisted dialect lexicography: Challenges and opportunities in processing historical Bavarian dialects. *Proceedings of the eLex 2025 Conference*, 453–475.

TEI Consortium. (2012). *TEI P5: Guidelines for electronic text encoding and interchange*.

Tiberius, C., Heylen, K., De Does, J., Vanroy, B., Vandeghinste, V., & Van Doeselaar, J. (2024). LLMs and evidence-based lexicography: Pilot studies at INT. In S. Krek (Ed.), *Book of abstracts of the workshop Large Language Models and Lexicography* (pp. 44–48).

Veronis, J., & Ide, N. (1991). An assessment of semantic information automatically extracted from machine readable dictionaries. In J. Kunze & D. Reimann (Eds.), *Fifth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 227–232). Association for Computational Linguistics. https://aclanthology.org/E91-1040/

Vulić, I., De Smet, W., & Moens, M.-F. (2012). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval, 16*(3), 331–368. https://doi.org/10.1007/s10791-012-9200-5

**Author Contributions: Emrah Özcan:** Conceptualization, Methodology, Investigation, Data interpretation, Funding acquisition, Project administration, Supervision. Writing-original draft: *Introduction, Theoritical Framework, Methodology, Structural modelling, Challenges in Multilinlingual Extraction, Future Directions, Conclusion*. **M. Fatih Erkoç:** Investigation, Methodology, Data interpretation, Project researcher. Writing-original draft: *Introduction, Theoritical Framework, Methodology, Structural modelling, Future Directions, Conclusion*. **Hasan Tokatlı:** Methodology, Data curation, Project researcher. Writing-original draft: *Introduction, Methodology, Conclusion*.

**Data Accessibility Statement:** The data that support the findings of this study are available from the corresponding author, [EÖ], upon request.

**Ethical Approval/Participant Consent:** There are no human participants in this article and informed consent is not required.

**Financial Support:** This study was supported by Yildiz Technical University, Scientific Research Projects Coordination Unit (General Research Project #SBA-2021-4256).