ItalicAI: A Conceptual Dictionary for the Italian Language Optimized for NLP and LLMs

Samuele Scuglia
Independent Researcher
samuele.scuglia@gmail.com

May 2025

1 Introduction

The Italian language remains relatively underrepresented in large-scale conceptual resources used for training artificial intelligence models, particularly in the domain of natural language processing (NLP) and large language models (LLMs).

While many multilingual LLMs provide token-level support for Italian, few public datasets offer clean, interpretable, and semantically aligned lexical structures.

We introduce **ItalicAI**, an open-source conceptual dictionary built from Italian linguistic resources, designed to offer atomic, non-overlapping semantic units. It is machine-usable, lossless in concept definition, and includes full morphological variation.

ItalicAI aims to support a wide range of applications, from tokenizer construction to semantic embedding alignment and transparent LLM reasoning.

2 Dataset Construction

The ItalicAI dataset was constructed by parsing the Italian edition of Wiktionary to extract clean groups of perfect synonyms. Only synonym clusters explicitly defined as equivalent, and bounded within semantic definition sections, were included. Ambiguous, partial, or nested synonym chains were discarded to preserve one-to-one conceptual alignment.

Each accepted group was converted into an atomic concept and assigned a unique identifier (‘CONC_XXXX‘). Subsequently, the dataset was enriched using **Morph-it**, a high-quality morphological lexicon for Italian. All inflected forms (verb conjugations, plural nouns, gendered adjectives, etc.) were matched to their corresponding base lemma and concept.

The result is a semantically disambiguated and morphologically complete lexicon, suitable for AI systems and LLM fine-tuning.

3 Format and Usage

The ItalicAI dataset is distributed in a modular and machine-friendly structure.

- **meta.pkl** – A NanoGPT-compatible mapping between token IDs and concept names (e.g., 'CONC_00213'), ready for direct model training or fine-tuning.
- **lista_forme_sinonimi.jsonl** – A JSONL file in which each line maps a single concept ID to:
  - its perfect synonyms (as listed in Wiktionary)
  - all inflected forms retrieved via Morph-it
- **lista_concetti.txt** – A plain text list of all concept IDs, one per line.

Each concept is unique and unambiguous, representing a single lexical meaning. All synonyms and forms are fully expanded, making the dataset suitable for:

- tokenizer development
- reverse lemmatization
- conceptual training of LLMs
- semantic-to-linguistic mapping

The structure is minimal, efficient, and designed to be loaded or indexed line-by-line for fast parsing in real-time NLP pipelines.

4 Future Work

The ItalicAI project was designed from the beginning as a modular and scalable system. While this release focuses on Italian, we aim to extend the approach to other languages—starting with English—and apply the same methodology: extracting clean, disambiguated synonym clusters and enriching them with complete inflectional forms.

Beyond multilingual expansion, ItalicAI is also the foundation for a new kind of interpretable and layered artificial intelligence. We are currently developing a lightweight LLM trained exclusively on this conceptual dictionary, with semantic tokens at its core and grammar handled by a separate module. This architecture promotes transparency and control over model behavior, and could support hybrid symbolic-neural reasoning in future AI systems.

Though secondary to this initial release, this training framework will become a key pillar of the broader ItalicAI vision.

5 Disclaimer and Collaboration

ItalicAI is the result of a one-man effort, developed independently outside of academic institutions, and built during nights and weekends after a full-time construction job as a waterproofing installer.

Given the scale and ambition of the project, minor inaccuracies or oversights may exist. All data has been carefully parsed, cleaned, and validated using programmatic tools, but the complexity of semantic grouping and morphological mapping leaves room for future refinement.

The project is fully open to:

- community feedback and suggestions
- external contributions or forks
- complete reengineering or repurposing

ItalicAI is not a finished product, it is a foundation. If you believe in conceptual modeling, multilingual AI, or the idea of traceable and interpretable tokenization, you are welcome to build on top of this work, or tear it apart and remake it better.

6 License

ItalicAI is released under the **Italica Open Knowledge and NonCommercial License (IOK-NC)**.

This license permits:

- Free use, analysis, and modification of the dataset for research, academic, educational, and non-commercial purposes.
- Forking and redistribution with proper attribution.

Commercial use, including integration into paid services, proprietary systems, or any monetized application, is strictly prohibited without prior written consent from the author.

To ensure legal protection and fairness, all disputes related to the license are to be resolved via arbitration administered by the **World Intellectual Property Organization (WIPO)**, based in Geneva, under the WIPO Arbitration Rules, in Italian.

This legal design ensures that ItalicAI remains available for public benefit and protected from exploitation or privatization by corporate entities.