Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study

Paola Marongiu, Simonetta Montemagni
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche (CNR-ILC)
Via Giuseppe Moruzzi 1, Pisa
{paola.marongiu, simonetta.montemagni}@ilc.cnr.it

Abstract

The increasing digital availability of dialectal atlases has significantly enhanced access to dialectal data and their potential for linguistic and cultural studies. However, despite their richness, such resources often remain difficult to integrate into contemporary data-driven research workflows, due to complex data structures and limited interoperability. Most digital dialectal atlases still rely on traditional access models centered on maps, offering only implicit and coarse-grained semantic structures, which limits concept-based exploration and potential for integration with other linguistic resources in the Linguistic Linked Open Data (LLOD) ecosystem. This paper presents a case study carried out on the *Atlante Lessicale Toscano* (ALT), aimed at addressing these limitations through the introduction of an explicit semantic layer designed to support both user-oriented exploration and machine-actionable interoperability. While ALT already provides a conceptual organization of dialectal materials, this structure was originally conceived for human navigation and not for integration with other computational lexical-semantic resources. To bridge this gap, we align ALT concepts with ItalWordNet, leveraging its synset-based model as a widely adopted semantic backbone in NLP and LLOD infrastructures. The case study focuses on the domain of agriculture, whose historically grounded conceptual distinctions are often underrepresented in general-purpose lexical resources. The paper proposes a mapping strategy, analyzes coverage and mismatch patterns, and releases a new aligned resource mapping ALT agricultural concepts to ItalWordNet, thereby creating the prerequisites for interoperability and reusability of dialectal atlas data.

Keywords: dialectal resources, WordNet, computational lexical semantics, specialized knowledge

1. Introduction

In recent years, a growing number of dialectal atlases have been made available in digital form, substantially enhancing access to and exploration of dialectal data. These resources play a key role in the preservation of cultural heritage and domain-specific knowledge. Beyond their linguistic relevance, dialectal atlases encode fine-grained conceptual distinctions related to traditional practices, environmental knowledge, and material culture, thereby representing a valuable source of cultural information. However, such resources often remain difficult to integrate into contemporary data-driven research workflows, due to complex data structures and limited interoperability.

We focus here on resources documenting dialects of Italy: notable examples include two digital versions of the *Sprach- und Sachatlas Italiens und der Südschweiz* (Atlante Italo-Svizzero, AIS, 1928-1940), namely *NavigAIS*[1] (Tisato, 2019) and *AIS Reloaded*[2] (Loporcaro et al., 2021), as well as atlases circumscribed to specific areas, such as the *Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects, ALD*[3] (Goebl, 2020) or the *Atlante Lessicale Toscano* (ALT)[4] (Picchi et al., 2001; Cucurullo et al., 2006). All of them are currently accessible online.[5] While their digitalization significantly enhances accessibility from a human-user perspective, it does not automatically guarantee compliance with FAIR principles (Wilkinson et al., 2016) or readiness for integration within the Linguistic Linked Open Data (LLOD) ecosystem.

Despite the growing availability of digital dialectal atlases, access to dialectal data largely mirrors the intrinsic organization of traditional linguistic atlases. To the best of our knowledge, with the partial exception of ALT (see below), none of the above mentioned resources currently provides a structured and machine-actionable semantic layer enabling concept-based queries. The primary access key typically remains the map, retrieved via a numerical identifier or title. As a consequence, dialectal data are findable and accessible in a broad sense, but only weakly interoperable and reusable in the sense defined by the FAIR principles. In particular, the absence of shared ontologies, standardized vocabularies, and Linked Data representations limits

[1] https://navigais-web.pd.istc.cnr.it/

[2] https://www.ais-reloaded.uzh.ch/

[3] https://www.ald.gwi.uni-muenchen.de/it/?db=ald1

[4] https://dbtvm1.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm

[5] More recent initiatives include the digitization of the *Atlante Linguistico Italiano, ALI*, a foundational resource for dialectological research on Italian dialects which, however, is still ongoing (Cerruti et al., 2025).

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 274–286
16 May 2026.

their integration with external datasets and reduces their potential for computational reuse.

Semantic organization in dialectal atlases is generally limited to broad semantic fields, which often correspond to individual atlas volumes. For instance, the AIS questionnaire was organized into eight parts, each including approximately 200 questions (therefore maps). While this coarse-grained and essentially flat organization may offer initial guidance within the atlas, it is not explicitly represented or operationalized in its digital counterparts. Moreover, such groupings tend to be highly heterogeneous: for example, AIS Volume 1 encompasses topics ranging from relationships, age, and love to birth, marriage, death, given names, body parts, body functions, and physical qualities and defects. This, in turn, hinders concept-driven exploration of the atlas and complicates its mapping to external semantic resources.

The picture sketched above highlights that the digital availability of dialectal atlases raises new challenges concerning semantic organization, conceptual accessibility, and interoperability, which represent core requirements for FAIR-compliant data. Addressing these challenges serves a twofold purpose: internally, it supports users in navigating dialectal data along explicit semantic paths; externally, it establishes the prerequisites for integration with other linguistic resources and tools. For dialectal resources to remain relevant and usable within this evolving ecosystem, they must become interoperable with computational infrastructures and widely adopted encoding and semantic standards. Against this background, this paper presents a case study carried out on the ALT dialectal resource, aimed at addressing these limitations through the introduction of an explicit semantic layer, designed both to enhance user-oriented exploration and to support machine-actionable interoperability.

It is worth noting that ALT already provides a conceptual organization of dialectal materials, articulated into macro-classes, micro-classes, and concepts corresponding to individual questions of the ALT questionnaire (Cucurullo et al., 2006). However, this organization was conceived primarily to facilitate access and navigation for human users and was not designed for integration with other computational resources. To bridge this gap, we decided to use WordNet (WN) — specifically ItalWordNet (IWN) (Roventini et al., 2016; Monachini et al., 2016). This choice is motivated by the widespread adoption of WN as a lexical-semantic backbone in NLP and LLOD infrastructures (Litta et al., 2025; Mambrini et al., 2021), its synset-based model enabling interoperable semantic mapping, and its suitability as a reference framework to assess the alignment between domain-specific dialectal concepts and general-purpose lexical knowledge. For our case study, we decided to focus on the domain of agriculture, whose meanings and conceptual distinctions reflect historically grounded categorizations that are often underrepresented or absent in contemporary general-purpose lexical resources.

Our contributions are threefold: (i) we define a strategy to tackle the methodological challenges involved in mapping domain-specific concepts to a general-purpose lexical-semantic resource; (ii) we provide a quantitative and qualitative analysis of concept coverage and mismatch patterns, highlighting gaps and differences in granularity between dialectal and computational lexical-semantic resources; and last but not least (iii) we release a new aligned resource linking ALT agricultural concepts to IWN synsets, thereby enhancing the interoperability of dialectal data within the LLOD ecosystem.

2. The *Atlante Lessicale Toscano*

The ALT is a specialized linguistic atlas aimed at documenting lexical variation across both diatopic and diastratic dimensions. Data were collected between 1974 and 1986 through a large-scale fieldwork campaign carried out by trained interviewers in 224 locations throughout Tuscany, a region of central Italy. A total of 2,193 informants were selected according to standard sociolinguistic criteria, including age, educational level, socio-economic status, and cultural background.

2.1. The Questionnaire

Data elicitation relied on a questionnaire comprising 745 items, primarily targeting variation in lexical choice, meaning, and pronunciation. Two main elicitation strategies were adopted. On the one hand, onomasiological questions started from a given concept and asked informants to provide the corresponding lexicalization, e.g. terms used to denote 'bread crumbs'. On the other hand, semasiological questions began with a dialectal form and required informants to specify its meaning or range of referents together with its pronunciation, e.g. the term *ceppo*, which in Tuscan varieties may denote a 'tree stump', a 'log', or a 'Christmas present'.

Answers to both types of questions were phonetically transcribed and enriched with informants' explanations, comments, and any additional linguistic material that emerged during the interviews, even when not directly related to a specific questionnaire item. Overall, the data collection process yielded several million individual responses, subsequently organized into approximately 380,000 structured records. Of these, about 350,000 correspond to canonical responses to questionnaire items attested in the surveyed locations - often accompanied by usage contexts and metalinguistic

remarks - while roughly 30,000 records document dialectal forms collected as supplementary material during fieldwork.

2.2. ALT Versions

ALT was first released in 2000 as a CD-ROM (Giacomelli et al., 2000), providing users with the possibility to explore dialectal data through complex, parameter-driven queries tailored to specific research needs. With the increasing availability of internet-based technologies, the original CD-ROM edition was eventually superseded by ALT-Web[6] (Montemagni et al., 2006), an online platform granting access to the full body of linguistic data collected for the atlas. This transition substantially expanded the potential user base, reaching not only dialectologists and linguists but also non-academic users interested in the linguistic and cultural heritage of Tuscany.

In the CD-ROM edition of the ALT, an attempt was made to address the difficulty of retrieving specific concepts from the questionnaire through a set of 338 keywords designed to identify thematic groupings of questions. Although this provided some support for navigating the questionnaire, it still required consulting the entire list. In ALT-Web, by contrast, an ontology-based search was introduced, representing a significant enhancement that substantially improved user accessibility.

ALT-Web should still be regarded as a legacy application, together with its associated data. A new version of the resource, named ALT-Web OPEN, is currently under development with the aim of complying with the FAIR principles (Cucurullo et al., 2025). The overarching objective of this initiative is to ensure the long-term preservation of ALT dialectal data, to improve sustainable and open access, and to enable interoperability with other linguistic resources and computational tools. The work presented in this paper is part of this process.

2.3. ALT-Web Conceptual Organization

One of the canonical access keys to the materials of a linguistic atlas collected through a questionnaire is question-based access: within the corpus, all attestations elicited in response to a selected question are retrieved. This type of access, however, presupposes familiarity with the questionnaire used for data collection, which is often articulated into hundreds of questions. As mentioned above, in the case of ALT, the data collection questionnaire includes 745 items.

[6] https://dbtvm1.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm

2.3.1. ALT-Web Ontology

To further support users of ALT-Web in identifying the question or set of questions relevant to their interests, an ontology was developed to organize the concepts investigated by ALT into a domain-based hierarchy. In particular, the set of concepts covered by the ALT questionnaire was organized into 13 macro-classes, derived from the original subdivision of the questionnaire into thematic sections, plus an additional miscellaneous class collecting questions not readily assignable to the identified macro-classes.

At the highest level of the ALT ontology are the macro-classes listed in the first column of Table 1. For each macro-class, a set of intermediate conceptual classes — hereafter referred to as micro-classes corresponding to more fine-grained semantic groupings — has been identified (reported in the second column of the table), yielding a total of 396 such associations. On average, each macro-class is articulated into 28 micro-classes. At a lower level of the ontology, each micro-class is associated with a set of low-level keywords (inherited from the CD-ROM version), which serve to define more specific semantic groupings within a micro-class. These are, in turn, linked either to (a) elementary concepts expressed by Italian lexical items, including both single words and multi-word expressions, corresponding to onomasiological questions (third column), or to (b) dialectal lexical items, corresponding to semasiological questions (fourth column). The terminal nodes of the ontology thus correspond to the questionnaire items themselves. The final column of Table 1 reports the number of items associated with each macro-class, for a total of 2,750 associations.

2.3.2. Organization Criteria

The criteria used to associate ALT questionnaire items with the conceptual classes of the ontology vary according to the type of question. Onomasiological questions are linked to the elementary concepts they investigate, which are in turn associated with progressively broader conceptual groupings up to the top-level nodes of the ontology. For onomasiological questions, each questionnaire item is classified under a single macro-class, while different semantic facets or possible cases of polysemy are captured through the association of the same question with multiple micro-classes. For example, question 169, aimed at collecting the lexicalizations for the concept 'head pad', is recorded in two different micro-classes under the macro-class AGRICULTURE, that of TOOL and that of TRANSPORTATION, highlighting different meaning facets (corresponding to "hyperonymy" and "used_for" relations).

In the case of semasiological questions, the pat-

Macro-class	Associated micro-classes	Onom. quest. items	Semas. quest. items	All quest. items
AGRICULTURE	39	124	220	344
FOOD AND NUTRITION	31	151	148	299
ANIMAL HUSBANDRY	16	67	87	154
WILD FAUNA	11	49	44	93
FORESTRY AND WOOD HARVESTING	23	60	145	205
HOUSING AND DOMESTIC ACTIVITIES	36	109	147	256
LANDFORMS AND GEOMORPHOLOGY	26	57	78	135
PLANTS AND FRUITS	23	115	60	175
CHRONOLOGICAL TIME	8	21	18	39
WEATHER AND METEOROLOGICAL PHENOMENA	17	64	35	99
HUMAN ACTIVITIES AND SOCIAL RELATIONS	32	68	90	158
HUMAN BEHAVIOUR AND EMOTIONS	70	84	316	400
HUMAN BODY AND CLOTHING	55	130	238	368
MISCELLANEOUS	9	7	18	25
Total	396	1106	1644	2750

Table 1: ALT-Web ontology: macro-classes and associated micro-classes and questionnaire items (onomasiological vs semasiological).

tern of associations at the terminal and pre-terminal levels of the ontology is completely different. Semasiological questions are connected to broader conceptual classes via the dialectal term under investigation. These associations were established based on the analysis of the meanings attested for each dialectal term in Tuscany. Therefore, a single semasiological question may be classified as relevant to multiple macro-classes. For instance, question 434d, aimed at eliciting the meanings of the dialectal term *manfano*, is associated — due to the marked polysemy of the term in Tuscan dialects — with the macro-classes AGRICULTURE, FORESTRY AND WOOD HARVESTING, PLANTS AND FRUITS, HUMAN BEHAVIOUR AND EMOTIONS, and HUMAN BODY AND CLOTHING. This distinction within the ALT ontology is shown in Figure 1 with reference to the semasiological question 434d *manfano*, whose attested meanings range across different macro-classes.

2.3.3. Ontology-based Search

This organization of the questionnaire enables concept-based access to dialectal materials. Users are first presented with the list of conceptual macro-classes; once a relevant macro-class has been selected, a set of micro-classes corresponding to more fine-grained conceptual groupings is displayed. By selecting a specific concept, users can then access the set of associated questionnaire items, further subdivided into two subsets according to question type. Figure 2 illustrates the results of a query targeting the micro-class TOOL, as displayed by the ALT-Web interface, in which questionnaire items are organized into two groups according to question type.

Unlike a flat list of lexical items, which merely provides keywords for retrieval, a conceptually structured resource offers an explicit network of relations among items, enabling more informed semantic interpretation. The ALT-Web ontology, therefore, functions as an effective, access-oriented semantic layer, designed to facilitate navigation and retrieval rather than to provide a logically complete or formally constrained conceptual model.

Notably, the same conceptual entity may appear under multiple macro-classes, reflecting alternative interpretative perspectives rather than a strict subsumption-based hierarchy. While this design choice enhances usability and supports exploratory access, it limits the degree of ontological systematicity that can be achieved, and has direct implications for the integration of external semantic resources. Our mapping onto IWN is intended to address this issue.

3. WordNet

The choice of WordNet (WN) as the target semantic resource for our mapping is not self-evident. The most appropriate strategy for linking dialectal atlas data with computational semantic resources remains an open issue. Existing approaches include the alignment of entries with Wikidata entities, as adopted in *VerbaAlpina* (Colcuc, 2020), or with Concepticon identifiers (List et al., 2016), as in the digitization of the *Tableaux Phonétiques* (Hans Geisler, 2021). In both cases, the linking strategies rely on flat inventories of entities or concepts associated with shared identifiers, rather than on an explicitly articulated conceptual hierarchy.

While such approaches effectively establish cross-resource correspondences, they do not provide a formal representation of conceptual organization or of the semantic relations connecting lexical meanings. For modeling dialectal atlas data, where lexical variation frequently reflects subtle semantic distinctions and patterns of polysemy, a

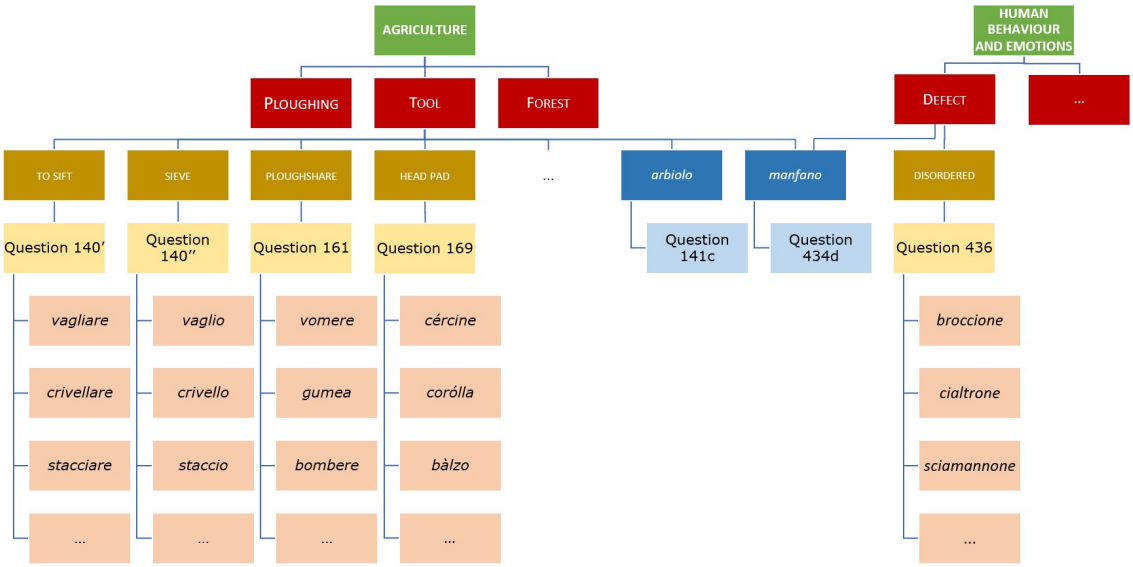

Figure 1: Fragment of ALT-Web ontology with onomasiological and semasiological questions, marked respectively by yellow and light blue boxes.

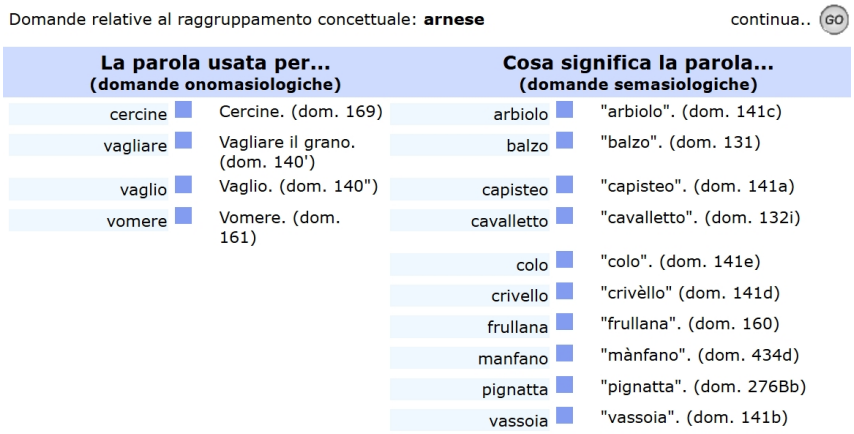

Figure 2: ALT-Web query results for the micro-class TOOL: onomasiological (left) and semasiological (right) questions. The keyword associated with each question is shown on the left: Question 140" "Vagliare il grano" 'to winnow grain' is associated to the keyword "vagliare" 'to winnow'.

structured lexical-semantic architecture appears more suitable. Moreover, in accordance with the FAIR principles, semantic interoperability requires the use of formal, shared, and machine-readable knowledge representations. A relational conceptual model provides a stronger foundation for both interoperability and reusability than simple identifier alignment.

Within this perspective, WN constitutes a particularly relevant modelling framework. Princeton WordNet (Fellbaum, 1998; Miller et al., 1990) was conceived as a lexical-semantic network in which meanings are represented not as isolated dictionary entries, but as nodes in an interconnected conceptual system. The fundamental unit of this architecture is the *synset* (synonym set), which corresponds to a single lexicalized concept. Each synset groups together lexical items (lemmas) that are interchangeable in a given context and thus express the same sense. WN draws a principled distinction between lemmas and senses: while a synset encodes one sense, a lemma may belong to multiple synsets. Polysemy is therefore modelled as the distribution of lemmas across distinct conceptual nodes, rather than as multiple definitions attached to a single lexical entry. For example, the English word *bank* appears in different synsets corresponding to a financial institution and to a river bank (Miller et al., 1990). Each synset is identified by a unique offset and is accompanied by (i) a gloss provid-

ing a definition of the concept, (ii) usage examples when available, and (iii) a set of formally typed semantic relations connecting it to other synsets. The most central relation is hypernymy/hyponymy, which structures the network hierarchically. Other relations include meronymy/holonymy (part–whole relations), as well as similarity and antonymy relations. This relational layer transforms WN into a structured semantic graph in which concepts are embedded within a formally defined hierarchy.

On the basis of the Princeton model, several language-specific WNs have been developed. Among them, Italian WordNet (IWN) (Roventini et al., 2016; Monachini et al., 2016) is a manually constructed lexical-semantic database for Italian, built in alignment with the original architecture while accounting for language-specific lexical and semantic distinctions. For instance, in IWN the synset with offset `#12355-n`, glossed as "mobile costituito da un piano sostenuto per lo più da quattro gambe" 'piece of furniture consisting of a flat surface typically supported by four legs', is associated with the lemma *tavolo* 'table' and is linked via hyperonymy to `#20176-n` "oggetto d'arredamento" 'furniture item', which in turn connects to more abstract nodes such as `#4716-n` "oggetto materiale, concreto" 'material, concrete object' and ultimately `#15544-n` "ciò che esiste" 'existing entity'.

Aligning dialectal concepts to WN is not intended to reduce their domain-specific richness. Rather, it anchors them within a shared and formally structured semantic space, thereby facilitating FAIR-oriented interoperability, cross-resource querying, and computational reuse.

4. Mapping ALT to IWN

Each level of the ALT ontology for the agricultural domain was mapped to a synset in IWN. The mapping process was carried out entirely manually, and involved all levels of the ALT-Web ontology for the agricultural domain. We started from the macro-class AGRICULTURE; then we mapped the micro-classes belonging under AGRICULTURE (39), e.g. PLOUGHING, TOOL, FOREST, etc.; finally, we mapped the concepts corresponding to onomasiological questions for each micro-class e.g., for the micro-class TOOL, VAGLIARE 'to winnow', VAGLIO 'sieve', VOMERE 'ploughshare', CERCINE 'head pad' etc. At this stage, semasiological questions have not been included in the mapping process, as the present study primarily aims to align the concepts investigated through the ALT questionnaire with IWN. The treatment of semasiological questions — namely, the association between dialectal items and their meanings as attested in Tuscany, already codified within the ALT ontology — will be deferred to a subsequent phase, contingent upon an assessment of the feasibility of mapping ALT concepts to IWN. The onomasiological questions for the macro-class AGRICULTURE are shown in Table A1 in Appendix.

In mapping a concept to a synset, we rely on the full text of the question that explicates the concept, rather than on the keyword alone (refer back to Figure 2). For instance, the keyword *grappolo* 'bunch' is associated with three distinct questions, and therefore with three different concepts: Question 115 GRAPPOLO D'UVA 'bunch of grapes'; Question 117 PARTE DEL GRAPPOLO D'UVA 'part of a bunch of grapes'; and Question 118 DOPPIO GRAPPOLO 'double bunch'. Each question is thus linked to a distinct synset, even when the keyword is identical. In order to identify the appropriate synset for mapping the concepts of the ALT ontology, we searched for corresponding concepts in version 2.0 of IWN (Monachini et al., 2016), querying both lemmas and glosses of the synsets.

The mapping process for such domain- and culturally specific concepts raised some challenges which required us to devise strategies to tackle them. These challenges can be grouped into different categories: i) extremely specific concepts that appear in ALT but do not have a corresponding synset in IWN (section 4.1) ii) multiple synsets available for the same concept (section 4.2) iii) ambiguous concepts that might be relevant to more than one synset (4.3).

4.1. Missing synsets for domain-specific concepts

In some cases, the dialectal material describes extremely fine-grained domain-specific concepts that do not have a directly corresponding synset available in IWN. This is the case for the concept VAGLIARE IL GRANO 'to winnow grain' (Question 140'), which is lexicalized in various ways in Tuscan dialects, e.g. *burattàre, conciàr l grano, nettàre l grane*. In Italian however, a specific lexeme to express this specific action is missing, and this gap is reflected by IWN. A blatant example of fine-grained concept distinctions is given by questions 115, 117 and 118, referring, respectively, to bunch of grapes, part of a bunch of grapes, and double bunch of grapes (see above), all three lexicalized in different ways across Tuscany. In order to keep track of such problematic cases, we opted for a two-level mapping: in the first level, we give the value "NA" to express the absence of a specific synset for that concept; in the second level, we map the concept onto the closest synset available in the IWN hierarchy. For instance, we map *vagliare il grano* on the second level onto a more generic synset `#38290-v` "vagliare" 'to examine'. Table 2 shows the two-level mapping for the concepts associated with the keyword *grappolo*.

It is worth mentioning that in some cases for an extremely specific concept an equally specific synset was found. That was the case for the concept RACCOGLIERE, which in ALT is paired with lexical entries describing the action of harvesting, inherently specific to agricultural activities. In this case, an extremely specific synset was found in IWN, paired with the lexical entry *raccogliere*, glossed as "prendere e radunare i frutti della terra" 'harvesting and gathering the fruits of the land'.

4.2. Multiple synsets available

In contrast, in some other cases, more than one synset could be mapped onto the same ALT concept. For instance, GRANO 'wheat' was relevant to three different synsets: `#26302-n` "pianta erbacea con foglie lineari e infiorescenza a spiga" 'herbaceous plant with linear leaves and a spike inflorescence'; `#26303-n` "frutto del grano da cui si ricavano farine alimentari; frutti cariossidi ('chicchi') da cui si ricava farina per pane e paste alimentari" 'fruit of wheat from which food flours are obtained; caryopsis fruits ('grains') from which flour for bread and pasta is produced'; `#8172-n` "pianta che produce cariossidi gialle" 'plant that produces yellow caryopses (grains)'. In such cases, we decided to map GRANO onto a primary synset, and to provide a second and third option of mapping to preserve coherence and ensure the possibility of exploiting the IWN hierarchy in all three cases. It should be noted that *grano* and *granoturco* (or *granturco*) 'corn' in IWN belong to the same synset (`#8172-n`), despite technically referring to two different types of grains (wheat vs. corn). In this case, we decided to preserve the IWN information as it is.

4.3. Conceptual collapse and polysemy in mapping

During the alignment process, we observed the presence of ambiguous concepts in the ALT hierarchy. A representative example are the questions investigating the concepts FRANTOIO and SECCATOIO DI CASTAGNE. In ALT, the concept FRANTOIO may refer either to the tool used to crush olives ('olive-press') or to the building where olive oil production takes place ('olive oil mill'). IWN distinguishes these meanings: the synset corresponding to *frantoio* refers to the machine used for crushing olives, while the building is described by a different synset (`#31669-n`), associated with the lemma *oleificio*. The ALT informants do not consistently separate these senses; rather, they reflect a conceptual blending of tool and location. For the time being, and in the absence of systematically disambiguated response forms, we mapped FRANTOIO to the synset describing the olive-oil ('macchina per la frangitura delle olive'), acknowledging that this choice only partially captures the semantic range observed in the data.

The case of SECCATOIO DI CASTAGNE presents a different challenge. Chestnut drying is an agricultural activity particularly common in the Tuscan-Emilian Apennine area. By inspecting the answers collected in ALT under the concept SECCATOIO, we found that, just as for FRANTOIO, the interviewees refer to a physical building (a dedicated construction or small house for chestnut drying), or to a specific tool or component used in the drying process (e.g., a stone disc or wooden apparatus). However, no direct synsets exist in IWN for any of these specific concepts. Several candidate hypernyms were considered for the second-level mapping described in section 4.1, in line with the internal structure of the ALT ontology: (i) a type of building (micro-class LUOGO 'place'), (ii) an instance of a drying process (ESSICCAZIONE), or (iii) an activity within chestnut cultivation (CASTAGNICOLTURA 'chestnut farming'). All three perspectives are semantically defensible. However, given the heterogeneity of the atlas responses, we opted to map the concept to the synset corresponding to *essiccazione* 'drying process', which captures the functional purpose common to all referents for this concept.

5. Results and analysis

We carried out an analysis of the results of our mapping on two axes. First, we measured to what extent the knowledge (i.e., the concepts) represented in ALT for the macro-class AGRICULTURE can be mapped onto a general-purpose computational semantic resource like WN (section 5.1). Then, we compared the knowledge organization in the ALT ontology (Figure 1) with the one extracted from IWN by leveraging the synsets relations (section 5.2).

5.1. Concept coverage

As described in Section 4, the mapping process raised different challenges due to the presence of very specific concepts in the ALT ontology. In order to understand to what extent a culture-specific resource like ALT can be mapped to a general-purpose resource like IWN, we measured the concept coverage by counting the number of concepts and micro-classes that we could be mapped to an IWN synset. The results are shown in Table 3. In order to give an account of the concept coverage, we decided to reflect the two-level mapping strategy explained in section 4.1 in our analyses. Out of 34 micro-classes, only for two of them we had to rely on a hypernym, using the second-level mapping. That was the case for CASTAGNICOLTURA 'chestnut farming' and PORZIONE DI FRUTTO 'fruit portion'. Concept mapping was more problematic:

Concept	Translation	Synset	Closest synset
GRAPPOLO D'UVA	bunch of grapes	#26969-n	–
PARTE DEL GRAPPOLO D'UVA	part of a bunch of grapes	NA	#26969-n
DOPPIO GRAPPOLO	double bunch (of grapes)	NA	#26969-n

Table 2: Two-level mapping for domain-specific concepts related to GRAPPOLO D'UVA 'bunch of grapes'.

some of them are very specific to Tuscan agricultural activities and artifacts, as described in section 4.1. Out of 38 concepts (i.e. questions), only 25 of them were mapped directly to their corresponding synset. The remaining 13 concepts were mapped on the second level to the closest synset found in IWN. Overall, 0.3% of the concepts and only 0.05% of the micro-classes did not find an exact correspondence with a synset. However, by using the two-level mapping, all concepts and micro-classes were successfully mapped to IWN.

5.2. Dialectal materials in the WordNet hierarchy

To explore the distribution of the ALT concepts in the IWN network, we went up the hypernyms chain defined within the WordNet structure, until we found the root synset for each mapping. Given our focus on the dialectal materials, we only worked on mapping of a synset to an ALT concept (in yellow in Figure 1), discarding micro-classes. In this section, we will i) give an overview of the concept specificity in the IWN network for all the concepts in the macro-class AGRICULTURE; ii) show an example of how the IWN hierarchy can support and/or clash with the handcrafted ALT ontology.

Concept specificity in the WN hierarchy. To determine whether the supposed cultural specificity of the concepts represented in ALT corresponds to a specificity of senses as they are represented in the WN network, we calculated the length of the hypernym chain for each mapping i.e., how many steps are required to go from the mapped synset to its root synset, that is, the most general one. The reasoning is that a synset positioned deep within the hierarchy (and therefore requiring a longer hypernym chain to reach the root synset) should correspond to a greater specificity of the concept described by the synset itself.

First, we only considered direct mappings i.e., concepts for which a synset was already available in IWN. The depth of each synset in the WordNet hierarchy per number of concepts associated is shown in Figure 3. Then, we included in the computation both first and second-level mappings i.e., concepts for which we had to rely on the closest available synset in the WN hierarchy due to the absence of a synset specific enough to describe the concept. The results are shown in Figure 4. For each second-level mapping we added an additional layer to make up for the missing synset.

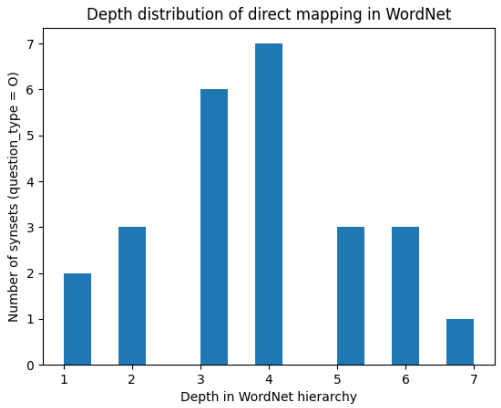

Figure 3: Synset depths for first-level mapping of concepts from ALT-Web for the macro-class AGRICULTURE.

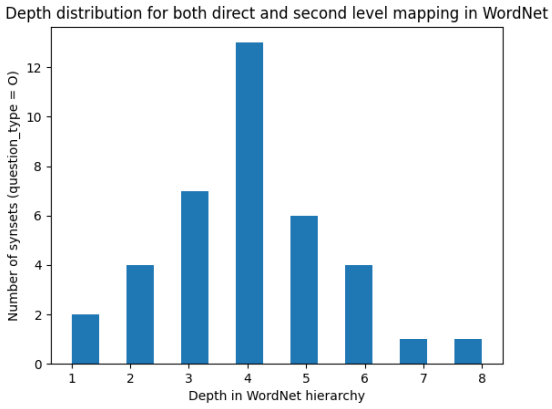

Figure 4: Synset depths for first and second-level mapping of concepts from ALT-Web for the macro-class AGRICULTURE.

In both cases, we find that the lexical layer of the ALT-Web ontology i.e., dialectal realizations in ALT-Web tends to align with more fine-grained WordNet nodes than the ontological key layer. As shown in Figures 3 and 4, most of the concepts (13 for first-level mapping only, and 20 for mapping at both levels) are located at the 3rd or 4th node of the semantic hierarchy in WordNet. A minority of concepts (7 for first-level mapping, and 10 for both levels) are located at the 5th, 6th, 7th, or event 8th node of the WordNet semantic hierarchy, suggest-

Level	Tot	Direct	Closest	% Closest
Micro-classes	34	32	2	0.05
Concepts	38	25	13	0.3

Table 3: Concept coverage for micro-classes and concepts in the agriculture macro-class (onomasiological questions only). Direct = concepts that were mapped directly to one synset; Closest = concepts that were mapped to the closest available synset.

ing that they refer to even more specific referents in the macro-class AGRICULTURE.

Concepts that appear deep in the WordNet hierarchy are related to words that describe specific activities in the field e.g. VAGLIARE IL GRANO 'to winnow grain' (offset `#38290-v`, 8th level of the hierarchy); MIETERE 'to reap' (offset `#32641-v`, level 6); TREBBIARE 'to thresh' (offset `#41163-v`, level 7). Concepts that appear at the 3rd and 4th level usually indicate similar activities, e.g. RACCOGLIERE LE OLIVE 'to harvest olives', probably less specific in the WN structure, or specific referents in the countryside: FASCIO DI SPIGHE LEGATE SUL CAMPO 'sheaf of ears of grain', PAGLIUZZA 'straw (small pieces)', RICCIO DELLA CASTAGNA 'chestnut burr'.

Handcrafted ontology vs. WordNet. We compared convergence and differentiation in IWN among the four concepts under the micro-class TOOL, shown in Figure 1: VAGLIARE IL GRANO 'to winnow grain', VOMERE 'ploughshare', VAGLIO 'sieve', CERCINE 'head pad'. The concepts VOMERE and VAGLIO display partially overlapping hypernymic trajectories, sharing higher-level synsets that correspond to general artifact and tool categories. This indicates that, despite referring to distinct implements, both items are semantically anchored in the same macro-domain of agricultural instruments. Their paths converge at abstract nodes encoding physical object and man-made artifact concepts, revealing a clear structural proximity in the WN hierarchy. By contrast, VAGLIARE IL GRANO, projects onto a different branch of the hierarchy, associated with action or process synsets rather than concrete artifacts. Its (second-level) mapping situates it within a procedural semantic field, thereby separating it structurally from the nominal tool-denoting items. This divergence reflects the ontological distinction between entities and events that is overlook in ALT. CERCINE follows yet another path. Although it ultimately converges with the other concepts at very abstract upper-level synsets (e.g., physical entity or artifact, depending on its mapping), its intermediate hypernyms differ, reflecting a more specialized or less centrally connected semantic niche within the agricultural ontology. Overall, we found that concepts belonging to the same thematic domain (agriculture) do not necessarily cluster at lower hierarchical levels. Instead, semantic convergence tends to emerge only at higher, more abstract synsets, while lower-level organization preserves fine-grained conceptual distinctions between tools, processes, and specialized objects.

It should be noted that, while WN provides a structured conceptual organization, its hierarchy does not necessarily reflect all possible perspectives on conceptual relatedness. Consider the concept ACINO 'grape'. IWN has a lexical entry for this word, assigned to two different synsets. One of the synsets is `#25757-n` "chicco dell'uva o di frutto simile" 'grape, or grain-like part of some other fruit'. Let us consider now the concept GRAPPOLO. This lemma has a lexical entry in IWN with which a few synsets are associated, again due to the polysemy of this word in Italian. The synset `#7317-n` "infruttescenza" 'infructescence' has the following hypernyms chain: infruttescenza > insieme dei frutti derivati da un'infiorescenza raggruppati in modo da sembrare un frutto unico 'a group of fruits derived from an inflorescence, clustered together so as to appear as a single fruit' > la totalità; persone o animali o cose radunate insieme 'the whole; people, animals, or things gathered together'. This is due to the fact that the prominent semantic feature of this word is not the notion of fruit, but the fact that it refers to a group of entities, working as a collective noun. For this reason, the user does not find in our mapping ACINO and GRAPPOLO D'UVA under the notion of fruit, but in two different paths of the IWN network.

6. Conclusion

This paper presents a case study on mapping the ALT to IWN, focusing on the agricultural domain. We discussed the manual mapping process, our methodological choices and associated challenges. By releasing this dataset, we show how a highly structured and rich dialectological resource such as ALT can be enhanced and enriched by adhering to the Open Data principles and making it interoperable with other computational resources. Building on this work, we made some preliminary considerations on how the ALT-WordNet dataset can be exploited to closer inspect the concepts related to agriculture, and make a tentative conceptual grouping based on the IWN architecture. Future work will focus on exploiting the ILI (Interlingual Index) codes in IWN (Bartolini and Quochi, to appear) to project the ALT concepts onto the English WordNet, and allow for cross-linguistic studies.

7. Author contributions

PM manually mapped the agricultural domain to Italian WordNet and performed the evaluation and analysis of results. PM wrote sections 3, 4, 5, 6, and the Appendix, and contributed to section 1. SM provided overall supervision throughout the mapping process and manuscript preparation, including review and editing, and wrote sections 1 and 2.

8. Acknowledgements

This work was supported by the following projects: the NRRP Project PE 0000020 “Cultural Heritage Active Innovation for Sustainable Society” - CHANGES, CUP B53C22003890006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU, within the activities of Spoke 3; the project “Una risorsa integrata del patrimonio linguistico-culturale toscano: l’Atlante Lessicale Toscano multimediale - multiALT” (’An Integrated Resource of Tuscan Linguistic and Cultural Heritage: the Multimedia Tuscan Lexical Atlas - multiALT’) funded by Regione Toscana (PROGETTI FSE+ 2021-2027) with the financial support of Accademia della Crusca.

9. Bibliographical References

Massimo Cerruti, Lorenzo Ferrarotti, Stefano Fiori, Matteo Rivoira, and Barbara Turchetta. 2025. Il progetto DigitALI: problemi e prospettive della digitalizzazione dell’Atlante Linguistico Italiano. *Linguistik online*, 137(5):37–67.

Beatrice Colcuc. 2020. La geolinguistica digitale e le sfide lessicografiche nell’era delle digital humanities: l’esempio di VerbaAlpina. In *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l’Informatica Umanistica*, pages 74–81, Milan, Italy. Associazione per l’Informatica Umanistica e la Cultura Digitale.

Nella Cucurullo, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi, and Eva Sassolini. 2006. Dialectal resources on-line: the ALT-web experience. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sebastiana Cucurullo, Simonetta Montemagni, Rubino Saccoccio, and Eva Sassolini. 2025. The Challenge of Obsolescence of Digital Archives in Cultural Heritage. A Case Study. *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, pages 506–512.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Hans Goebl. 2020. Presentazione delle due parti (ALD-I e ALD-II) dell’Atlante linguistico del ladino dolomitico e dei dialetti limitrofi. *Romance Philology*, 74:245–265.

Johann Mattis List Hans Geisler, Robert Forkel. 2021. A digital, retro-standardized edition of the tableaux phonétiques des patois suisses romands (tppsr). In André Thibaut, Matthieu Avanzi, Nicolas Lo Vecchio, and Alice Millour, editors, *Nouveaux regards sur la variation dialectale*, pages 13—-36. Editions de Linguistique et de Philologie, Strasbourg.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Eleonora Litta, Marco Carlo Passarotti, Valerio Basile, Cristina Bosco, Andrea Di Fabio, and Paolo Brasolin. 2025. Liita: a knowledge base of interoperable resources for italian. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 130–135.

Michele Loporcaro, Stephan Schmid, Chiara Zanini, Diego Pescarini, Giulia Donzelli, Stefano Negrinelli, and Graziano Tisato. 2021. Ais, reloaded: A digital dialect atlas of italy and southern switzerland. In André Thibaut, Matthieu Avanzi, Nicolas Lo Vecchio, and Alice Millour, editors, *Nouveaux regards sur la variation dialectale*, pages 111—-136. Editions de Linguistique et de Philologie, Strasbourg.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking valency frames and wordnet synsets in the lila knowledge base of linguistic resources for latin. In *Further with Knowledge Graphs*, pages 16–28. IOS Press.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Eugenio Picchi, Simonetta Montemagni, and Lisa Biagini. 2001. DBT-ALT: A System for Storing and Querying the Data of the Atlante Lessicale

Toscano (ALT). *Dialectologia et Geolinguistica (DiG)*, 9:85–103.

Graziano Tisato. 2019. Acquisizione dell'intero AIS (Sprach- und Sachatlas Italiens und der Südschweiz). In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale*, number 6 in Studi AISV, pages 131–153. Officinaventuno.

Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

10. Language Resource References

Roberto Bartolini and Valeria Quochi. to appear. ItalWordNet OMW. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Giacomelli, Gabriella and Agostiniani, Luciano and Bellucci, Patrizia and Giannelli, Luciano and Montemagni, Simonetta and Nesi, Annalisa and Paoli, Matilde and Poggi Salani, Teresa. 2000. *DBT-ALT: Atlante Lessicale Toscano in CD-rom*. Lexis Progetti Editoriali.

Monica Monachini, Claudia Soria, and Antonio Toral. 2016. ItalWordNet kyoto. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Montemagni, Simonetta and Picchi, Eugenio and Cucurullo, Sebastiana and Paoli, Matilde and Sassolini, Eva. 2006. *ALT-WEB: l'Atlante Lessicale Toscano in rete*. Istituto di Linguistica Computazionale "Antonio Zampolli", CNR.

Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Resources

We release the ALT-WordNet dataset and associated scripts in the following GitHub repository `https://github.com/paoma370/ALT-Web-to-WordNet.git`.

Appendix

Question Id	Question text (IT)	Question text (ENG)
87	Bosco di castagni	Chestnut grove
88	Bosco di castagni giovani	Young chestnut grove
89	Pollone di castagno	Chestnut shoot
90	Terriccio di castagno	Chestnut soil
91	Riccio della castagna	Chestnut burr
92	Buccia interna della castagna	Inner skin of the chestnut
93	Castagne vuote contenute nel riccio	Empty chestnuts in the burr
94	Seccatoio di castagne	Chestnut drying house / tool
115	Grappolo d'uva	Bunch of grapes
117	Parte del grappolo d'uva	Part of a bunch of grapes
118	Doppio grappolo	Double bunch (of grapes)
119	Raspo, grappolo senza acini	Grape stalk (bunch without grapes)
121	Viticcio	Tendril (of the vine)
122	Chicco d' uva	Grape
124	Sostegno morto della vite	Stake for the vine
125	Filare	Row (of vines)
126	Raccogliere le olive	To harvest olives
127	Frantoio	Olive oil mill / olive-press
128	Segale	Rye
129'	Mietere	To reap
129"	Mietitura	Reaping / harvest
130	Fascio di spighe legate sul campo	Sheaf (of ears of grain)
133	Palo del pagliaio	Haystack pole
134	Trebbiare	To thresh
135	Resti del grano nel campo	Remains of grain in the field (stubble)
135a	Resti del granturco nel campo	Remains of maize in the filed (corn stubble)
137	Pula	Chaff
139	Pagliuzza	Straw (small pieces)
140'	Vagliare il grano	To winnow grain
140"	Vaglio	Sieve
161	Vomere	Ploghshare
162	Bigoncia	Large wooden tub (for grapes/wine)
163	Grossa cesta rotonda per l'erba	Large round basket for grass
165c	Come si chiama il cestino per le olive?	What is the basket for olives called?
167	Orcio per l'olio	Oil jar
169	Cercine	Head pad (for carrying loads)
517	Scegliere	To select, to sort
519B	Raccogliere	To gather, to harvest

Table A1: Questions in the agriculture macro-class.

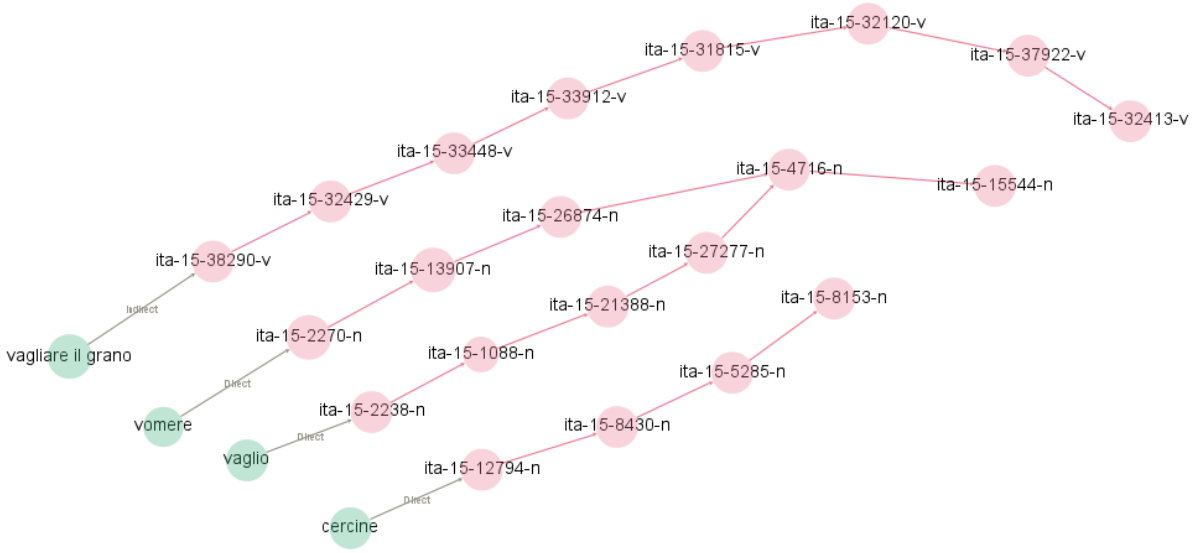

Figure A1: Combined WordNet subgraph for the concepts VAGLIARE IL GRANO, VOMERE, VAGLIO, and CERCINE, from the micro-class TOOL, for comparison with Figure 1. Green nodes represent the ALT concepts, while pink nodes represent IWN synsets. Coloured edges represent hierarchical relations between synsets (hypernymy). First- and second-level mapping are represented by grey edges, respectively indicated by the labels 'Direct' and 'Indirect'. As explained in section 5.2 in the main text, edges labeled as 'Indirect' are counted as one additional hierarchical step in the analysis. The concept described by the synsets in Figure is shown in Table A2.

synset_id	gloss (IT)	gloss (EN)
ita-15-1088-n	arnese atto all'esecuzione di determinate operazioni proprie di un'arte, di un mestiere o di una tecnica	tool used to perform specific operations typical of an art, craft, or technique
ita-15-12794-n	panno avvolto sulla testa per appoggiarvi brocche o ceste	cloth wrapped around the head to support jugs or baskets
ita-15-13907-n	piastra, lastra molto sottile	plate, very thin sheet
ita-15-15544-n	ciò che esiste	that which exists; entity
ita-15-21388-n	arnese o dispositivo atto a compiere determinate operazioni	tool or device designed to perform specific operations
ita-15-2238-n	attrezzo per setacciare	tool for sifting
ita-15-2270-n	lama dell'aratro che rovescia la zolla tagliata	plough blade that turns over the cut sod
ita-15-26874-n	pezzo di materiale rigido, di poco spessore e in genere rettangolare	piece of rigid material, thin and generally rectangular
ita-15-27277-n	tutto ciò che gli uomini producono	everything produced by human beings
ita-15-4716-n	oggetto materiale, concreto	material, concrete object
ita-15-5285-n	una parte di un tutto	a part of a whole
ita-15-8153-n	ciascuna delle porzioni o degli elementi in cui è diviso un tutto	each of the portions or elements into which a whole is divided
ita-15-8430-n	pezzo di tessuto	piece of fabric

Table A2: Synsets with Italian gloss and English translation from Figure A1.