Reading Between the Clues: A Fine-Grained Analysis of Italian Crossword Solving Systems

Cristiano Ciaccio[1,2,†], Asya Zanollo[3,4,†], Alessio Miaschi[2], Gabriele Sarti[5], Felice Dell'Orletta[2] and Malvina Nissim[6]

[1]*Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo 3, Pisa, Italy*
[2]*Institute for Computational Linguistics "A. Zampolli" (CNR-ILC) - ItaliaNLP Lab, Via G. Moruzzi 1, Pisa, Italy*
[3]*University School for Advanced Studies IUSS Pavia, Piazza della Vittoria 15, 27100, Pavia, Italy*
[4]*Laboratory for Neurocognition, Epistemology, and Theoretical Syntax - NeTS-IUSS, Piazza della Vittoria 15, Pavia, Italy*
[5]*Khoury College of Computer Sciences, Northeastern University, 440 Huntington Avenue, 202 West Village H, Boston, USA*
[6]*Center for Language and Cognition (CLCG), University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands*

Abstract

Existing evaluations of crossword solving systems predominantly rely on aggregate metrics, offering limited insight into how different systems behave with respect to the linguistic properties of clues and answers. In this work, we propose a fine-grained evaluation on the Italian language along two complementary annotation axes: a *syntactic* axis, characterizing each clue according to an improved categorization of the *cruciverbese* register, and a *lexical-semantic* axis, capturing the clue-answer relation through cosine distance, named entities, and degree of polysemy. We release an enriched version of an existing Italian crossword dataset annotated along both axes, introduce two novel architectures for crossword clue answering, and conduct an extensive comparative analysis of a broad set of existing systems together with our newly proposed models, surfacing systematic differences in how architectural choices interact with the linguistic properties of clue-answer pairs.

Keywords

NLP, Crossword Solving, Evaluation, Language Models, Italian

1. Introduction and Background

Crosswords (CWs) offer a unique way to engage with language. Their resolution requires sensitivity to ambiguity, lexical retrieval, attention to orthography, commonsense and cultural knowledge, and some degree of inferential reasoning, especially over the most compact and cryptic definitions. Indeed, the language of CWs is notably idiosyncratic: clues operate through deliberate misdirection, exploitation of polysemy, and, more generally, genre-specific conventions that usually diverge rather sharply from ordinary discourse. Such features make CW definitions a particularly fertile testing ground for human cognitive skills and Neural Language Models (NLMs).

From an evaluation and systems development point of view, crossword puzzles have long been investigated within the NLP community as a particularly challenging testbed: their resolution requires not only linguistic competence but also extensive world knowledge, cultural awareness, and lateral thinking skills [1, 2, 3, 4]. Traditional approaches to crossword solving relied on retrieval-based methods and shallow lexical and semantic features to identify relevant information [5, 6]. With the rise of NLMs, the focus has shifted toward systems capable of capturing more nuanced inferential relations although their effectiveness remains constrained by fundamental limitations in accessing linguistic and culturally-relevant knowledge, especially for less-resourced non-English languages [7]. For the Italian language, in particular, crossword solving has been a long-standing research interest, evolving from early clues-to-clues retrieval-based methods [8, 9] – typically struggling with clues requiring non-literal

CLiC-it 2026: Twelfth Italian Conference on Computational Linguistics, September 14–16, 2026, Palermo, Italy
[†]These authors contributed equally.
cristiano.ciaccio@ilc.cnr.it (C. Ciaccio); asya.zanollo@iusspavia.it (A. Zanollo); alessio.miaschi@ilc.cnr.it (A. Miaschi); g.sarti@northeastern.edu (G. Sarti); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); m.nissim@rug.nl (M. Nissim)
0009-0001-6113-4761 (C. Ciaccio); 0009-0001-3987-4843 (A. Zanollo); 0000-0002-0736-5411 (A. Miaschi); 0000-0001-8715-2987 (G. Sarti); 0000-0003-3454-9387 (F. Dell'Orletta); 0000-0001-5289-0971 (M. Nissim)


reasoning and wordplay – to more contemporary neural approaches. Recently, Ciaccio et al. [10] introduced CROSSWORD SPACE, a collection of asymmetric dual-encoder architectures, demonstrating that learning a shared latent space between clues and answers outperforms traditional retrieval-based baselines and generalizes to neologisms. This growing interest led to Cruciverb-IT [11], the first shared task on Italian crossword solving, held at EVALITA 2026 [12]. Cruciverb-IT comprised two subtasks – answering individual clues given a target answer length, and autonomously solving complete crossword grids – resulting in a total of 17 system submitted, with the best systems combining retrieval-augmented LLMs, fine-tuned encoder-decoders, and constraint-based grid solvers [13, 14].

Crossword solving is also a recognized tool for cognitive exercise in educational and medical settings [15, 16, 17, 18] fostering research on automated CW generation and linguistic analysis of the *cruciverbese* register [19]. Unlike general-purpose CWs, where clues may rely on obscure trivia or slang, educational puzzles are designed to reinforce specific learning objectives and cognitive skills. It is within this pedagogical framework that the necessity for a dedicated linguistic analysis of crossword language becomes apparent to ensure that clues are not only semantically accurate but also structurally varied and syntactically sound. To this end, Zeinalipour et al. [20] proposed a syntactic classification of CW clues highlighting a considerable and valuable syntactic variability in Italian CW clues.

Despite the growing body of work on crossword solving in NLP, a systematic study of how clue and answer linguistic properties influence system behavior is still missing. Existing benchmarks and shared tasks predominantly rely on aggregate metrics such as accuracy and Mean Reciprocal Rank (MRR), offering limited insights into why a given system succeeds or fails on specific portions of the data. Understanding how solvers interact with different types of clues – for instance, syntactically complex constructions, clues involving named entities, or highly polysemous answers – is essential not only to characterize the strengths and weaknesses of current approaches, but also to inform the development of more targeted models and resources, including, for example, controllable crossword generation systems that explicitly account for stylistic and structural variability. This strongly relates to the usage of CWs in educational and geriatric environments, allowing the autonomous generation of puzzles which linguistically meet the solver's needs, by creating more semantically straightforward clues with a simple syntax, or by reinforcing linguistic traits and skills that tend to decline with cognitive ageing.

In this work, we address this gap by proposing an extensive evaluation of crossword solving systems for Italian along two complementary annotation axes: a *syntactic* axis, which classifies each clue according to an improved categorization scheme of the *cruciverbese* register, and a *lexico-semantic* axis, which captures the clue-answer relation divergence through the cosine distance between their embeddings, the presence and type of named entities, and the number of senses associated with the answer. Building on this framework, our contributions are threefold: (i) we release an enriched version of the dataset introduced by Zeinalipour et al. [19] and adopted in Cruciverb-IT [11], annotated along both axes; (ii) we propose CROSSWORDSPACE++, a novel system for clue answering built upon the CROSSWORDSPACE architecture of Ciaccio et al. [10], which achieves competitive performance on the Cruciverb-IT task; (iii) we conduct an extensive comparative analysis of existing Italian crossword solving systems and our newly developed model along the proposed annotation axes, revealing systematic differences in how solvers handle distinct linguistic and semantic configurations of clue-answer pairs[1].

2. Approach

In the following sections, we firstly describe the approach leveraged to build our new CROSSWORDSPACE++ system and, later, describe the process we used to annotate the dataset along the two axes of complementary linguistic dimensions, i.e., the syntactic and lexical-semantic categorization.

[1]The resources are available at the following repository: https://github.com/snizio/crosswordspacepp.

2.1. CrosswordSpace++

Our approach to clue-answering is structured as an extension to the system proposed and developed in [10], CrosswordSpace, a transformer-based dual-encoder that learns a shared latent space between clues and solutions. Although dual-encoders are highly efficient and can retrieve relevant candidate pools that often contain the correct answer, these architectures struggle to produce high-precision rankings because both inputs are processed independently by different encoders, resulting in static representations of both clues and answers. For example, the word *bat* may answer the clues *An anagram of TAB*, *Dracula's alter ego*, or *A piece of pine*, and a dual encoder must compress all of these meanings into a single point in the shared space. Crucially, in the context of crosswords, the meanings of clues and answers are deliberately ambiguous, requiring the joint processing of clues and answers to exploit as much information as possible and fully resolve the ambiguity.

To address this, we adopt a *retrieve and re-rank* strategy [21, 22], in which (a) the CrosswordSpace dual-encoder acts as a fast, high-recall retriever, returning a ranked list of candidate solutions with retrieval scores; (b) a cross-encoder – which jointly encodes clue and candidate and can therefore exploit cross-token dependencies inaccessible to the dual-encoder – rescores those candidates, producing a second set of scores; and (c) the two score sets are linearly interpolated to yield the final ranking. This fusion preserves the retriever's global ranking signal while letting the cross-encoder correct local ordering errors, exploiting the complementarity of the two scoring functions.

2.2. Dataset

For our experiments, we rely on the dataset introduced in Zeinalipour et al. [19], further extended in Ciaccio et al. [10] using the official splits released in Ciaccio et al. [11] consisting of a training (90%), validation (5%) and a test (5%) set. The resulting collection contains approximately 410,000 clue-answer pairs, covering a wide range of puzzle types, including wordplay, cryptic clues, named-entity initials, and fill-in-the-blank clues. This choice ensures a consistent and fair comparison across all systems considered in our analysis: both the models submitted to the EVALITA shared task and the systems we developed are trained on the same training set, while the held-out test set serves as the basis for the analyses presented in the following sections.

2.3. Models

Our study considers two complementary sets of systems. The first set consists of the dual encoder originally introduced in [10], together with three novel extensions, CrosswordSpace++ whose architectures are described in detail in Sec. 3.1. The second set comprises the systems submitted to the Cruciverb-IT shared task at EVALITA 2026, for which we consider the best-performing run of each participating team. Specifically, we include: **AC/DG** [23], a retrieval-based system combining BM25 and dense retrieval; **FFT-UniBa** [14], an IT5-based encoder-decoder fine-tuned with length-aware special tokens; **MINDS** [24], an encoder-only approach framing clue answering as masked language modeling over variable-length spans; **UNIBA** [25], an IT5-Large encoder-decoder trained to generate answers conditioned on partial solutions; and **UniTor** [13], a retrieval-grounded system relying on structured LLM prompting with GLM 4.6. We refer the reader to the EVALITA overview paper [11] and to the individual system reports for a comprehensive description of each approach.

2.4. Syntactic Categorization

In the previous works on *cruciverbese*, Zeinalipour et al. [20] propose a rule-based syntactic categorization of Italian CW clues using spaCy's dependency parser [26]. The categorization is organized into macro- and micro-categories, reflecting a hierarchical annotation structure. Macro-categories are identified by examining the root node of each clue and capture their primary syntactic distinction between nominal and clausal structure. Micro-categories are derived by gradually subdividing macro-categories through the combination of additional binary features, such as the presence of an

Category	Clue	Answer	Cos↓	NER	Senses
PP	Tra duo e quartetto	*trio*	0.243	—	1
act:clitic	Li ha il cobra indiano ma non quello africano	*occhiali*	1.020	—	2
act:missSubj	Circondano Saturno	*anelli*	0.764	LOC	22
act:pron	A lui spetta la sentenza	*giudicenaturale*	0.277	—	0
adjP	Gustose	*saporite*	0.141	—	1
bare_NP	Strisce di tessuto elastiche	*fasce*	0.625	—	21
bare_NP:rel	Argomento che non può mancare negli oroscopi	*amore*	0.802	—	17
biclausal	Si dice logori chi non ce l'ha	*potere*	1.003	—	8
cop:clitic	Lo è il canto di Mameli	*inno*	0.321	PER	2
cop:missSubj	Sono uguali in Spagna	*aa*	1.015	LOC	0
cop:pron	Sono prestigiose quelle di Oxford e Cambridge	*università*	0.470	ORG	2
def_DP	Il grande eroe di Burgos	*cid*	1.023	LOC	0
def_DP:rel	L'attività di chi gareggia	*agonistica*	0.326	—	1
impersonal_reflexive	Si dice di casi rari, isolati	*sporadici*	0.273	—	2
ind_DP	Un tipo di barca a vela	*star*	0.994	—	1
ind_DP:rel	Un militare che vola	*aviatore*	0.365	—	1
inf_VP	Alimentarsi	*nutrirsi*	0.138	—	1
passive	Viene chiamata Maestà	*regina*	0.268	MISC	4

Table 1
Representative examples per syntactic category (CAT), selected to span the extremes of cosine distance between clue and answer embeddings (Cos↓), named-entity type (NER; — if none), and number of word senses of the answer. Only categories with ≥100 examples are shown. The unknown category is excluded.

article, a relative clause (RCl), a copular verb etc. For instance, within the nominal macro-category, the combination of article=True and definite=True yields a definite Determiner Phrase (DP) clue (see [20] for a comprehensive description of the full categorization framework). In the current work we present an ameliorated syntactic categorization of CW clues in three different dataset. Specifically, *metalinguistic* clues - previously defined by the length of the answer - have been analyzed with respect to their syntactic structures and included in their corresponding category. Three new categories have been added: **biclausal** - including clues like *Chi vi entra cerca un titolo, libreria* ('Those who enter there are looking for a title, bookstore') which are composed of a main and a subordinate clause with both inflected verbs; **relativeCl:isolated** identifying clues having a relative pronoun as syntactic root node, such as *Che si riferisce a un tipo di cannocchiale astronomico, telescopico* ('That refers to a type of astronomical telescope, telescopic'); and **ind_DP:rel** for clues with an indefinite DP as nominal head and a RCl as modifier, like *Un ruminante che vive fra i dirupi delle montagne, camoscio* ('A ruminant that lives among mountain crags, chamois'). Clues with passive verbs, previously categorized as *pass:missSubj* or *pass:other*, distinguished on the basis of the omitted argument, have been included in the same **passive** category to avoid wrong categorization due to parsing errors. The same has been done for clues with impersonal or reflexive verbs, now categorized uniquely as **impersonal_reflexive**. The category *advP* has been deleted because it involved cases in which the adverb was present in the structure without being the root node. Clues previously categorized as *bare_Det* have been included in other categories for a more fine-grained annotation.
To assess the reliability of the automatic categorization, we manually validated a stratified sample of 100 clue-answer pairs drawn from the test set, sampled in proportion to the frequency of each category. The annotation was blind. Overall agreement between manual and automatic categorization is around 80% (see Table 4 in the Appendix B). The updated categories, the annotation results and the error analysis are described in detail in the Appendix B.

2.5. Lexico-Semantic Categorization

In parallel with the syntactic categorization described in Section 2.4, we enrich the dataset with a second layer of annotations aimed at characterizing the lexical and semantic properties of each clue-answer pair. This layer is organized along three complementary dimensions: the semantic distance between clue and answer, the presence and type of named entities, and the degree of polysemy of the answer. A representative sample of the dataset annotated with this second layer, along with syntactic

categorization, is reported in Table 1.

Cosine Distance. For each clue-answer pair, we compute the cosine distance[2] between the embedding of the clue and the embedding of the answer, both obtained through `sentence-bert-base-italian-xxl-uncased`[3], an Italian encoder specifically fine-tuned as a sentence-transformer. Stemming from the literature on automatically assessing the creativity of texts [27], this measure provides a proxy for the semantic and distributional divergence between the question and its solution: clues whose lexical content is closely related to the answer (e.g., paraphrastic or definitional clues) are expected to exhibit lower distances, whereas clues relying on indirect associations, wordplay, or world knowledge should yield higher distances expressing divergent ideas that are unlikely to be paired together. Cosine distance offers a continuous indicator of how directly an answer can be recovered from the surface form of its clue, thereby implicitly gathering a signal about the degree of creativity – in terms of convergence/divergence – of the relationship between clue and answer.

Named Entities. To assess the role of referential and encyclopedic content in crossword solving, we run a Named Entity Recognition (NER) system on the concatenation of clue and answer for each pair in the dataset. We leveraged `bert-italian-finetuned-ner`[4], an Italian encoder fine-tuned to predict token-level entity types: PER (person), LOC (location), ORG (organization), and MISC (miscellaneous). For each clue-answer pair, we record the full sequence of token-level tags and, separately, the dominant entity type on the clue side and on the answer side, defined as the most frequent label within each segment. The extracted entities allow us to distinguish between pairs that contain named entities, typically associated with more specific, anecdotal, or culturally grounded knowledge, and pairs that do not, which tend to rely on more general lexical and conceptual associations.

Word Senses. Finally, we annotate each answer with its number of senses, used as a proxy for lexical ambiguity and, indirectly, for the complexity of the retrieval task. To this end, we rely on the Italian Wiktionary[5]: each answer's surface form is first looked up directly; if no entry is found, we fall back to a stem-based match using the NLTK Snowball Italian stemmer[6] – the answer is reduced to its stem and matched against the stems of all dictionary entries, with ties broken by edit distance to the original answer. From the resulting entry, we extract the full list of senses across all part-of-speech, excluding inflected forms, and take their count as the polysemy value. Answers of three characters or fewer are excluded from this annotation, since the stem-based fallback becomes unreliable at that length.

3. Experimental Setting

3.1. CrosswordSpace++

The dual-encoder retriever, CROSSWORDSPACE, is fine-tuned following the procedure and hyperparameters of Ciaccio et al. [10][7] on top of the `paraphrase-multilingual-mpnet-base-v2`[8] multilingual sentence encoder (≈278M parameters per branch, ≈556M total).

The cross-encoder is trained as a binary classifier over clue-answer pairs. Given a clue c, a positive answer s_{pos}, and a negative answer s_{neg} – both of equal character length – the model is fine-tuned to assign label 1 to the input $x_{\text{pos}} = c \oplus s_{\text{pos}}$ and label 0 to $x_{\text{neg}} = c \oplus s_{\text{neg}}$, where $\oplus$ denotes sequence concatenation with a separator [SEP] token. Negatives are mined from the candidate pool retrieved by the already-trained dual-encoder, stratified by retrieval cosine score relative to the positive's score p:

[2] Given the representations of a clue $\mathbf{c}$ and a gold answer $\mathbf{s}$, the cosine distance d is computed as $d(\mathbf{c}, \mathbf{s}) = 1 - \frac{\mathbf{c} \cdot \mathbf{s}}{||\mathbf{c}|| \times ||\mathbf{s}||}$

[3] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased.

[4] https://huggingface.co/nickprock/bert-italian-finetuned-ner.

[5] We used the parsed version introduced in Ciaccio et al. [28].

[6] https://www.nltk.org/api/nltk.stem.snowball.html.

[7] We refer the reader to the original paper for the technical details.

[8] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

three *hard* negatives sampled from candidates with scores in $[0.65,\ \min(0.85,\ p-0.05)]$, three *medium* negatives from $[0.5,\ 0.65)$, and three *easy* negatives sampled randomly from the remaining tail. The margin on the hard band, set as -0.05 to the positive's score, was set to avoid drawing negatives from the immediate neighborhood of the gold answer. To further reduce the risk of false negatives, we discard any candidate that (i) appears as an alternative gold answer for the same clue in the augmented training data, (ii) shares the stem with the gold, or (iii) is listed as a synonym – or shares a stem with a synonym – of the gold in the Italian Wiktionary. This procedure yields nine negatives per positive, expanding the training set by a factor of ten.

We experiment with three backbones for the cross-encoder: `bert-base-italian-xxl-cased`[9] (≈110M parameters), `sentence-bert-base-italian-xxl-uncased`[10] (≈110M parameters) and `multi-sentence-BERTino`[11] (≈67M parameters). All models are trained for up to ten epochs with a binary cross-entropy loss with logits and a positive-class weight of nine to counterbalance the 1:9 positive-to-negative ratio, using AdamW with learning rate 2×10^{-5}, weight decay 0.01, a linear warmup over the first 5% of training steps and a batch size of 256. The validation MRR@10 is evaluated every 5% of training steps, and the best checkpoint is retained.

At inference, the retriever R, given a clue c and a vocabulary $W = \{w_1, \ldots, w_n\}$, returns a ranked list of the top $k = 1000$ lenght-filtered candidates $S = (s_1, \ldots, s_k)$ with cosine-similarity scores $\{r_i\}_{i=1}^{k}$. The cross-encoder CE then rescores the top $m = 100$ candidates,[12] producing scores $\{c_i\}_{i=1}^{m}$. The final ranking combines the two via linear interpolation,

$$\text{score}(c, s_i) \ = \ \alpha\, c_i + (1 - \alpha)\, r_i, \tag{1}$$

with $\alpha \in [0, 1]$ tuned on the validation set by grid search over $\{0.00, 0.01, \ldots, 1.00\}$, selecting the value that maximizes MRR@10. The full pipeline contains ≈666M parameters with the BERT cross-encoder and ≈623M with BERTino.

3.2. Evaluation

In terms of evaluation metrics we employ the same ones used within Cruciverb-IT: Acc@1/10, that is the accuracy in retrieving the correct solution word given the corresponding clue, considering the top 1 and 10 candidates produced by the systems; and Mean Reciprocal Rank (MRR@10), that is the average of the reciprocal ranks of the first relevant item across the top 10 retrieved answers.

To assess differences and correlations between system performance and the two annotation axes, we complement these metrics with a set of statistical analyses. For the lexico-semantic axis, we apply the Mann–Whitney U test [29] to evaluate whether the cosine distance between clue and answer embeddings differs significantly between correctly retrieved answers (*hits*) and incorrect ones (*misses*), at two cutoffs corresponding to top-1 and top-10 retrieval. As an effect size measure, we report the rank-biserial correlation. The coefficient ranges from -1 to $+1$, with positive values indicating that hits tend to have lower cosine distance than misses — i.e., semantically closer clue-answer pairs are easier to retrieve — and negative values indicating the opposite. The analysis is conducted both globally and separately for each syntactic category, restricted to those containing at least 100 samples to ensure statistical reliability.

For the analyses by NER entity type and by number of senses, we report hit rates (Acc@1 and Acc@10) directly for each group: NER entity types occurring in the clue (O, LOC, ORG, PER, MISC) and sense-count bins of the answer (0, 1, 2–3, 4–6, 7+), respectively. These values are descriptive rather than inferential, as the primary goal of this analysis is to characterize how the presence of named entities in the clue and the lexical ambiguity of the answer relate to retrieval difficulty across systems.

[9] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased.
[10] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased.
[11] https://huggingface.co/nickprock/multi-sentence-BERTino.
[12] Since $m < k$ and the cross-encoder operates only within the retrieved set, the top-m accuracy of the full pipeline is upper-bounded by the Acc@m of the retriever.

4. Results

In the following, we present the results of our analysis: we first characterize the dataset along the two proposed annotation axes (Section 4.1), then report the performance of our novel architectures against the Cruciverb-IT systems (Section 4.2), and finally provide a fine-grained comparative analysis along the syntactic (Section 4.3) and lexical-semantic (Section 4.4) axes.

4.1. Linguistic Profile of the Dataset

The relevant distributions — syntactic category frequencies, cosine distance, named entities, and word senses — are reported in Fig. 3 (Appendix A). The distribution of **syntactic categories** in the present datasets aligns with the structural patterns observed in the ItaCW corpus [20]. Simple nominal clues – bare NP and definite DP – represent the most frequent structures, followed by active clauses missing subject and indefinite DP. This structural distribution confirms prior findings indicating that while CW language exhibits high syntactic variability, it remains fundamentally driven by nominal structures.

Cosine distance reveals distinct traits in different syntactic constructions. Infinitive VP display the highest semantic consistency (lowest cosine distance) to their answers, demonstrating a highly literal and predictable structural environment. Conversely, PP, copular clause missing subject, and biclausal clues have the highest median distances, suggesting these clausal and prepositional frameworks are structurally leveraged to construct indirect, trivia-reliant riddles. In the case of PPs, their high distance and variance are a direct byproduct of their inherent syntactic versatility - capable of modifying nouns or verbs to denote time, location, or instrumentality - which maps them onto highly divergent vector representations. Most standard nominal structures like bare_NP and def_DP, have very similar median distances: the semantic addition of a uniqueness condition via the definite article ([30]) does not drastically change the overall predictability of the answer. Similarly, appending a RCl modifier to an NP head yields minimal variance: bare_NP:rel displays a minimal increase in distance, whereas definite and indefinite DPs modified by a relative clause exhibit the opposite trend. For most grammatical structures, clues are consistently designed to maintain a steady level of indirectness or wordplay.

Across nearly every linguistic category, the median number of **word senses** is remarkably low, sitting firmly between 1 and 2. This demonstrates that regardless of clue complexity, CW targets overwhelmingly favor concrete, specific lemmas over highly ambiguous ones. However, distinct structural variations emerge. Standard noun phrases (bare_NP, def_DP) and certain verbal structures (act:missSubj, impersonal_reflexive) uniquely stretch to target words containing 10 to 25 distinct senses. Syntactically, these simple nominal structures represent highly flexible vehicles for contextualizing heavily polysemous words. In contrast, PP constructions exhibit the lowest, most compressed sense distribution, with a median close to 1 and virtually no high-reaching outliers. Compared to cosine distance results, when a clue is constructed as a PP, while being highly indirect (as evidenced by its high cosine distance), it leads to a very unambiguous, single-sense answer.

Finally, the vast majority of clues do not contain any **named entity**. However, the distribution is markedly uneven: definite DPs (alone and with relative-clause modifiers), copular and active clauses with pronominal or clitic subject, and bare NPs feature the highest proportion of NEs in the clue, while infinitive VPs, biclausal clues, impersonal-reflexive constructions, and adjectival phrases remain almost entirely entity-free. This is tied to the semantic function of these structures within a definitional context such as CWs: definite DPs and copular clauses, which inherently serve to isolate or predicate over a unique entity, naturally recruit proper names as anchors to the answer; conversely, infinitive and impersonal constructions, which lack a referential anchor in their syntactic core, tend to convey more abstract, generic content. Among entity types, PER is by far the most frequent across the dataset, followed by LOC, while ORG and MISC remain marginal across all categories.

Model	Acc@1	Acc@10	MRR@10
CrosswordSpace	0.58	0.81	0.66
C.E. SBertino	0.48	0.79	0.58
C.E. Bert	0.54	0.82	0.63
C.E. SBert	0.58	0.83	0.68
Cross++ SBertino (α = 0.16)	0.65	0.85	0.72
Cross++ Bert (α=0.18)	0.66	0.85	0.73
Cross++ Sbert (α=0.17)	0.68	0.85	0.74
UniTor	0.69	0.83	0.74

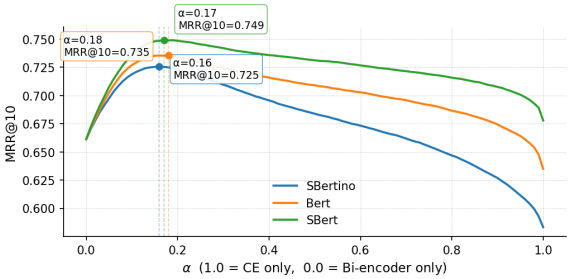

Figure 1: Left: test set performance of our systems and the best performing Cruciverb-IT system UniTor. **Right**: performance on the validation set as a function of α, where $\alpha = 0$ corresponds to dual-encoder performance only and $\alpha = 1$ corresponds to cross-encoder performance only.

4.2. CrosswordSpace++ Results

Figure 1 (left) reports the performance of our systems – both in terms of individual components and the full pipeline – on the Cruciverb-IT official test set. Every CrosswordSpace++ (Cross++) configuration outperforms the CrosswordSpace (Cross) dual-encoder by a substantial margin, with MRR gains of up to $+8$ points, confirming the effectiveness of the retrieve-and-rerank design. While the cross-encoder alone (C.E. in the Table) underperforms or only matches the dual-encoder, interestingly, the fusion strategy always yields higher results, suggesting that the two components have different but complementary capabilities. In other words, the cross-encoder is not a stronger ranker than the dual-encoder, but a different one. This interpretation is supported by the behaviour of MRR as a function of the scoring weight α. As shown in Figure 1 (right), MRR rises sharply for small α and peaks in the range $0.16 \leq \alpha \leq 0.18$ across all cross-encoder backbones, decaying monotonically for larger values. The optimal solution assigns roughly four-fifths of the weight to the retriever and only one-fifth to the reranker, providing gains only when its score distribution is strongly skewed towards a specifically high confidence solution that the dual-encoder alone was not able to identify.

Qualitatively inspecting the cases where Cross++ gains the most over the dual-encoder alone shows two main patterns recurring. The first concerns clues on specific encyclopedic or referential knowledge: e.g., for *"Giacomo poeta"* the dual-encoder returns generic surnames (*bandello*, *carducci*, *altoviti*, …) while the fusion ranks *leopardi* in the top ten candidates. In such cases, the dual-encoder correctly identifies the semantic kind of the answer but fails to captures specific details in the clue. The second pattern involves clues exploiting wordplay or metonymy, where the surface semantics of the clue actively mislead the retriever: for *"Una sosta per gli internauti"* → *sito* the dual-encoder retrieves some broad and general meaning of *sosta* (*coma*, *cala*, with *sito* at rank 88), and only the joint encoding recovers the intended answer at rank 2; an analogous behaviour is observed for the wordplay clues *"Quello giudiziario non interessa enigmisti!"* → *casellario* and *"Il disk del computer"* → *hard*. These observations also explain why the full pipeline yields substantially similar scores across cross-encoder backbones (Cross++ Sbertino follows Cross++ Sbert behind by only 2 points MRR) that, on the other hand, report different performances: the two components attend to fundamentally different aspects of clues underlining the efficacy of the retrieve-and-rerank approach. Finally, our best configuration (Cross++ Sbert) matches the MRR of the top-ranked system submitted to the Cruciverb-IT challenge and surpasses it in Acc@10, despite using a backbone roughly two orders of magnitude smaller than the GLM-4.6 (≈357B parameters) employed by Unitor.

4.3. Syntactic Categorization Results

Table 2 shows the MRR scores for each system, including ours, in function of the syntactic categories discussed in Section 2.4. Categories are sorted by the row-wise average across all systems (the **Avg.** column), which provides a system-independent proxy for the intrinsic difficulty of each category. First of all, the row-wise average is closely followed by most systems: categories with high average MRR (*def_DP*, *act:pron*, *ind_DP:rel*, *passive*) tend to be the easiest for nearly every model, while low-average

	Ours: Cross	Ours: C.E.	Ours: Cross++	Unitor	FFT	MINDS	AC-DG	UNIBA	Avg.	Count
def_DP	.7	.7	$.77_{(+.07)}$	**.85**	.68	.69	.67	*.51*	.70	3304
act:pron	.62	.74	$.73_{(+.11)}$	**.78**	.67	.66	*.59*	.67	.68	104
ind_DP:rel	.74	.67	$\mathbf{.8}_{(+.06)}$	.74	.71	.65	.68	*.43*	.68	105
passive	.7	.69	$\mathbf{.78}_{(+.08)}$	.75	.68	.63	.63	*.55*	.68	227
def_DP:rel	.66	.69	$.73_{(+.07)}$	**.8**	.64	.64	.63	*.52*	.66	371
adjP	.71	.73	$\mathbf{.79}_{(+.08)}$	.71	.68	.66	*.51*	.54	.67	973
bare_NP	.68	.68	$.73_{(+.05)}$	**.75**	.65	.64	.59	*.48*	.65	5974
act:clitic	.64	.7	$\mathbf{.74}_{(+.1)}$	**.74**	.65	.63	.59	*.5*	.65	249
bare_NP:rel	.65	.67	$\mathbf{.73}_{(+.08)}$	**.73**	.62	.65	.61	*.51*	.65	242
impersonal_reflexive	.63	.69	$\mathbf{.73}_{(+.1)}$	.69	.64	.63	.6	*.49*	.64	1023
unknown	.67	.69	$\mathbf{.75}_{(+.08)}$	.71	.65	.63	.57	*.48*	.64	683
act:missSubj	.65	.68	$\mathbf{.73}_{(+.08)}$	.72	.64	.62	.57	*.49*	.64	2229
inf_VP	.63	.72	$\mathbf{.74}_{(+.11)}$	.67	.58	.55	*.5*	.56	.62	496
cop:pron	.59	.64	$.68_{(+.09)}$	**.7**	.57	.61	.57	*.55*	.61	223
ind_DP	.62	.63	$.69_{(+.07)}$	**.7**	.59	.58	.59	*.43*	.60	2006
biclausal	.57	.66	$\mathbf{.7}_{(+.13)}$	.63	.58	.58	.53	*.49*	.59	602
cop:missSubj	.6	.6	$.66_{(+.06)}$	**.68**	.59	.59	.57	*.42*	.59	593
PP	.63	.53	$.73_{(+.1)}$	**.74**	.59	.58	.53	*.33*	.58	384
cop:clitic	.51	.61	$.62_{(+.11)}$	**.64**	.52	.49	*.43*	.47	.54	936
ρ (L.O.O.)	0.74^*	0.64^*	0.7^*	0.77^*	0.87^*	0.88^*	0.61^*	—		

Table 2
MRR across all systems and syntactic category sorted by descending row wise average MRR (Avg. column). Categories with less than 100 examples are discarded.

categories (*cop:clitic*, *PP*, *cop:missSubj*, *biclausal*) are uniformly harder. To further assess the impact of the syntactic category on performance, we compute, for each system, the Spearman rank correlation between its per-category MRR scores and the leave-one-out average MRR across the remaining systems, measuring how closely each system's performance tracks the cross-system consensus. As shown in Table 2 (last row ρ L.O.O.), most systems exhibit a strong and statistically significant positive correlation with the consensus ranking, ranging from $\rho = 0.61$ for AC-DG to $\rho = 0.89$ for MINDS (all $p < 0.01$). These correlations support the fact that the syntactic category of the clue effectively captures structural properties that modulate the difficulty of clue-answering. A possible syntactic explanation of these results regards the presence of constituent movement in the structure. On the one hand, structural transformation is absent in the easiest categories, which consist of a well-formed phrase or sentence like: def_DP, *La filosofia dell'essere, ontologia* ('The philosophy of being, ontology'), act:pron, *Quella d'aria semina distruzione, tromba* (lit. 'That of air sows destruction, trumpet'), ind_DP:rel, *Un vizio che può dar noia al prossimo, fumo* ('A vice that bothers others, smoke') and passive *Viene trasmesso con le ultime informazioni, notiziario* ('It is transmitted with the latest information, news bulletin'). In such cases the answer is a synonym - in nominal clues - or the subject of the clause, which is simply omitted from its sentence-initial position. On the other hand, harder clues like cop:clitic involves clitic insertion and postverbal subject as in *Lo sono le mire dell'ambizioso, alte* (lit. 'So are the aims of the ambitious one, high'). For what concerns PP clues like *Nel ginocchio, menisco* ('In the knee, meniscus') , in standard language they are nested in bigger phrases, so their present in isolation can make the interpretation less straightforward, impacting their resolution, as previously said in 4.1. Finally, biclausal clues like *Ha la pelle di chi è emozionato, oca* (lit. 'It has the skin of one who is emotionally moved, goose') represent structurally complex constructions, composed of two clauses with two inflected verbs and they mainly target idiomatic expressions or proverbial sayings, which characterized them as more costly in terms of linguistic processing and semantically cryptic. The only exception is the cop:missSubj category like *È il voto peggiore, zero* (lit. 'It is the worst grade, zero'). Despite the insertion of a copula, these clues exhibit no structural movement and generally behave like nominal clues, which theoretically should make them straightforward to resolve.

Two systems deviate visibly from this trend: **UNIBA** ranks last or near-last in almost every category (which, in fact, shows no statistically significant correlation in ρ L.O.O.), with particularly sharp drops on *PP* (.33), *ind_DP* (.43), and *ind_DP:rel* (.43) – categories where the other systems remain competitive;

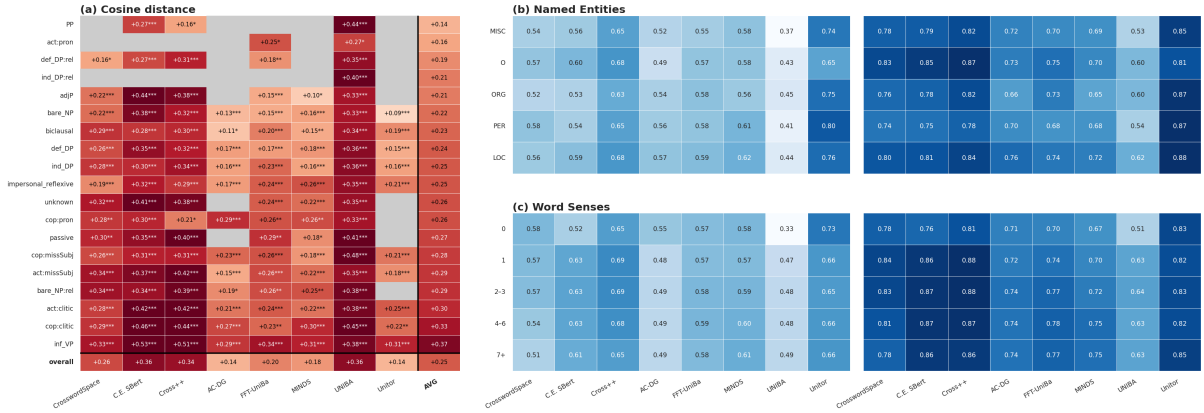

Figure 2: **(a)** Rank-biserial correlation (r) from Mann-Whitney U tests comparing cosine distance between correctly (hits) and incorrectly retrieved (misses) answers at Acc@10, broken down by syntactic category and system. Grey cells denote non-significant results ($p \geq 0.05$). The **AVG** column reports the mean r across systems; the **overall** row reports the result of a single test across all categories combined. Rows are sorted by the AVG column. **(b)** Acc@1 and Acc@10 broken down by NER entity type present in the clue (O = no entity). **(c)** Acc@1 and Acc@10 broken down by the number of word senses of the answer.

and **AC-DG**, which underperforms the average on *adjP* (.51), *inf_VP* (.50), and *cop:clitic* (.43).

Although we achieved the same MRR as **Unitor**, our Cross++ system either matches or exceeds it on a non-trivial subset of categories: *adjP* (.79 vs. .71), *ind_DP:rel* (.80 vs. .74), *passive* (.78 vs. .75), *biclausal* (.70 vs. .63), *inf_VP* (.74 vs. .67), *act:missSubj* (.73 vs. .72), and *bare_NP:rel* (.73, tied). The subscripted values in the Cross++ column denote the absolute MRR gain over the dual-encoder retriever alone (Cross). These gains range from $+.05$ on the simplest, most lexically transparent category (*bare_NP*) to $+.13$ on *biclausal* clues, and exceed $+.10$ on *act:pron* ($+.11$), *inf_VP* ($+.11$), *cop:clitic* ($+.11$), *act:clitic* ($+.10$), *impersonal_reflexive* ($+.10$), and *PP* ($+.10$). These difference suggests that, as previously observed, the cross-encoder contribution is concentrated in categories where the clue exhibits richer syntactic structure – biclausal constructions, clitics, PP, impersonal or reflexive verbs – and is smallest where the clue is a bare nominal description, which is the most widespread and straightforward in CWs. This is in line with the characteristics of the cross-encoder, which jointly encodes clue and candidate and can therefore exploit cross-token dependencies that the dual-encoder cannot by construction.

4.4. Lexico-Semantic Categorization Results

Cosine Distance. Figure 2 (a) reports the rank-biserial correlation from Mann-Whitney U tests comparing cosine distance between correctly (hits) and incorrectly (misses) retrieved answers at Acc@10, broken down by syntactic category and system. Across systems and categories, the coefficients are predominantly positive, indicating that hits systematically exhibit lower clue-answer cosine distance than misses – cosine distance is thus a meaningful behavioural and clue-answer divergence axis along which model errors organize. The strength of this effect is category-dependent. Categories previously identified as having higher median semantic distance and greater linguistic indirectness (PP, cop:missSubj, and biclausal clues) show the strongest and most consistently significant coefficients. The same pattern emerges in structures involving clitic displacement (cop:clitic, act:clitic), where the syntactic dislocation of an argument widens the semantic gap between clue and answer affecting retrieval difficulty. To assess whether this co-variation between cosine effect size and category difficulty is systematic across systems, we compute, for each system, the Spearman rank correlation between its per-category MRR and the per-category rank biserial effect size. All systems except UNIBA exhibit a strong and significant negative correlation (see Appendix C for the overall correlation results), ρ ranging from -0.63 for Unitor to -0.8 for MINDS, all $p < 0.005$, confirming that the categories in which cosine distance most strongly discriminates correct from incorrect predictions are precisely the categories in which models perform worst. The convergence between these two analyses importantly suggests that

syntactic complexity and high clue-answer semantic distance co-vary as complementary expressions of the clue indirectness that characterizes the *cruciverbese* register, strengthening both annotation axes.

Named Entities. Figure 2 (b) reports Acc@1 and Acc@10 of all evaluated systems broken down by the type of named entity occurring in the clue (O, LOC, ORG, PER, MISC). Our models — CrosswordSpace, C.E. SBert, and Cross++ — achieve their highest top-10 scores on the O class (e.g., Cross++ reaches Acc@10 of .87 on O, against .78–.84 on entity-bearing categories), suggesting that these retrieval-based architectures are particularly effective at modelling clues grounded in general lexical knowledge, wordplay, and definitional reasoning, which align well with the inferential relations captured by their contrastive objective. The opposite trend is observed for Unitor, which obtains its weakest top-10 score on O (.81) and its strongest on entity-bearing clues, peaking on LOC (.88), PER (.87), and ORG (.87). This is consistent with the nature of the backbone: relying on a large instruction-tuned LLM (GLM 4.6), Unitor benefits from the encyclopedic knowledge accumulated during pre-training when the clue revolves around specific entities. Accordingly, the gap between Unitor and the other systems is more pronounced on entity-based categories than on O, indicating that the comparative strength of LLM-based approaches is largely concentrated on clues whose resolution involves world knowledge associated with named entities. The remaining systems show a more uniform behaviour across categories, with no marked specialization toward either entity-bearing or entity-free clues, while UNIBA displays the lowest scores overall, with a particularly pronounced drop on MISC and PER.

Word Senses. Figure 2 (c) shows for CrosswordSpace a clear monotonic decrease as polysemy increases, with Acc@10 dropping from .84 on monosemous answers (bin 1) to .78 on highly polysemous ones (bin 7+), and Acc@1 declining from .57 to .51. This is consistent with an inherent limitation of dual encoder architectures, where clue and answer are encoded independently making alignment with the specific sense evoked by the clue more difficult. In contrast, both C.E. SBert and Cross++ exhibit a more robust behaviour across the 5 bins, with the former remaining essentially flat (Acc@10 within .86–.87) and the latter maintaining its peak up to the mid-polysemy range (Acc@10 of .88 on bins 1 and 2–3, with a mild decrease to .86 on 7+). This robustness can be attributed to the joint encoding of clue and answer performed by the cross-encoder, which can dynamically disambiguate among the possible senses of the answer by directly attending to the lexical and semantic cues provided by the clue — offering concrete empirical support for the design choice motivating Cross++. Unitor displays a distinctive pattern, with a sharp spike in Acc@1 on the 0 bin (.73) — which collects answers without any Wiktionary entry, typically proper nouns, neologisms, or specialized terminology — again likely reflecting the broader lexical coverage of its underlying LLM. For the remaining bins, it stays relatively stable (Acc@10 around .82–.85), while FFT-UniBa, MINDS, and AC-DG all show a mild positive trend across the polysemy spectrum, and UNIBA remains the weakest system across all bins.

5. Discussion and Conclusion

We presented a fine-grained evaluation of Italian crossword solving systems, based on two complementary annotation axes: a *syntactic* axis, capturing the structural properties of clues through an improved categorization scheme of the *cruciverbese* register, and a *lexico-semantic* axis, characterizing the clue-answer relation through cosine distance, named entities, and degree of polysemy of the answer. Using an enriched version of the dataset introduced in Cruciverb-IT [11] and two architectures extending the Crossword Space dual-encoder of Ciaccio et al. [10], we compared existing and newly developed systems along both axes. Besides the effectiveness of the ensemble strategy, with Cross++ achieving competitive performance while requiring a substantially smaller parameter footprint, our results show that both syntactic and semantic features affect model performance. In particular, the performance variance across different syntactic categories indicates that syntax plays an active role in clue resolution.

The kind of fine-grained evaluation proposed in this work opens several directions for future research. First, while our analysis focused on clue answering in isolation, the proposed categorization could provide valuable insights when extended to the full grid-solving setting, where the interaction between clue-level difficulty and structural constraints may reveal further patterns in system behaviour. Second,

the syntactic and lexico-semantic categorization could be leveraged not only as an evaluation lens, but also as a control signal for the automatic generation of crossword grids, enabling the construction of puzzles calibrated to different levels of complexity and stylistic variability. Finally, another promising direction is the extension of the proposed evaluation framework to human crossword solving: comparing the resolution patterns of human solvers and automated systems along the same annotation axes could shed light on similarities and divergences between human and machine strategies, and offer a more theoretically grounded basis for the design of cognitive-based approaches in educational and medical settings, where crosswords have long been employed as tools for linguistic and cognitive exercise.

Acknowledgments

This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. Partial support was also received by the project "Ordering Data for Efficient Representation Learning" (IsCd8_ORDER), funded by CINECA under the ISCRA initiative, for the availability of HPC resources and support.

Declaration on Generative AI

During the preparation of this work, the author used Claude and Gemini to generate images and to conduct grammar and spell checking. The author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] E. Wallace, N. Tomlin, A. Xu, K. Yang, E. Pathak, M. Ginsberg, D. Klein, Automated crossword solving, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3073–3085. URL: https://aclanthology.org/2022.acl-long.219. doi:10.18653/v1/2022.acl-long.219.

[2] J. Rozner, C. Potts, K. Mahowald, Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 11409–11421. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/5f1d3986fae10ed2994d14ecd89892d7-Paper.pdf.

[3] S. Saha, S. Chakraborty, S. Saha, U. Garain, Language models are crossword solvers, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 2074–2090. URL: https://aclanthology.org/2025.naacl-long.104/.

[4] A. Sadallah, D. Kotova, E. Kochmar, What makes cryptic crosswords challenging for LLMs?, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 5102–5114. URL: https://aclanthology.org/2025.coling-main.342/.

[5] G. Angelini, M. Ernandes, M. Gori, Webcrow: A web-based crosswords solver, in: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 295–298.

[6] G. Angelini, M. Ernandes, M. Gori, Solving italian crosswords using the web, in: International Conference of the Italian Association for Artificial Intelligence, 2005. URL: https://link.springer.com/chapter/10.1007/11558590_40.

[7] G. Sarti, T. Caselli, M. Nissim, A. Bisazza, Non verbis, sed rebus: Large language models are weak solvers of Italian rebuses, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 888–897. URL: https://aclanthology.org/2024.clicit-1.96/.

[8] G. Barlacchi, M. Nicosia, A. Moschitti, A retrieval model for automatic resolutionof crossword puzzles in italian language, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa, Pisa University Press, 2014, pp. 33–37.

[9] A. Moschitti, M. Nicosia, G. Barlacchi, SACRY: Syntax-based automatic crossword puzzle resolution sYstem, in: H.-H. Chen, K. Markert (Eds.), Proceedings of ACL-IJCNLP 2015 System Demonstrations, Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Beijing, China, 2015, pp. 79–84. URL: https://aclanthology.org/P15-4014/. doi:10.3115/v1/P15-4014.

[10] C. Ciaccio, G. Sarti, A. Miaschi, F. Dell'Orletta, Crossword space: Latent manifold learning for italian crosswords and beyond, in: Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[11] C. Ciaccio, G. Sarti, A. Miaschi, F. Dell'Orletta, M. Nissim, Cruciverb-IT @ EVALITA 2026: Overview of the Crossword Solving in Italian Task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[12] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[13] A. Shcherbakov, D. Croce, R. Basili, Unitor at cruciverb-it: Retrieval-augmented two-step reasoning for italian crossword clue answering, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[14] A. Porcelli, F. Di Gravina, E. Fontana, M. Curri, F. D. Di Gregorio, Fft-uniba at cruciverb-it: Special length tokens and csp for italian crossword solving, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[15] W. Orawiwatnakul, Crossword puzzles as a learning tool for vocabulary development, Electronic Journal of Research in Education Psychology 11 (2013) 413–428.

[16] E. Yuriev, B. Capuano, J. L. Short, Crossword puzzles for chemistry education: learning goals beyond vocabulary, Chemistry education research and practice 17 (2016) 532–554.

[17] D. P. Devanand, Computerized games versus crosswords training in mild cognitive impairment: The cog-it trial, Alzheimer's & Dementia 19 (2023) e073094.

[18] L. A. Wang, T. E. Goldberg, P. D. Harvey, A. J. Hanson, J. Motter, H. Andrews, M. Qian, R. Zhang, M. Janis, P. M. Doraiswamy, et al., Crossword puzzle training and neuroplasticity in mild cognitive impairment (cogit-2): 78-week, multi-site, randomized controlled trial with cognitive, functional, imaging and biomarker outcomes, International journal of clinical trials 12 (2025) 111.

[19] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Italian crossword generator: Enhancing education through interactive word puzzles, in: Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), 2023, pp. 455–464.

[20] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Italian crossword generator: an in-depth linguistic analysis in educational word puzzles, IJCoL. Italian Journal of Computational Linguistics 11 (2025).

[21] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).

[22] A. Yates, R. Nogueira, J. Lin, Pretrained transformers for text ranking: Bert and beyond, in: Proceedings of the 14th ACM International Conference on web search and data mining, 2021, pp.

1154–1156.
[23] A. Yassine, C. Savelli, D. Napolitano, G. Gallipoli, L. Cagliero, E. Baralis, Ac/dg at cruciverb-it: Retrieval-based approaches for italian crossword clue answering, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[24] F. Giobergia, Minds at cruciverb-it: Solving italian crossword clues with masked language models and candidate pooling, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[25] P. Basile, Uniba at cruciverb-it: Solving crosswords with encoder-decoder models and beam search, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.
[26] I. Montani, M. Honnibal, M. Honnibal, A. Boyd, S. Van Landeghem, H. Peters, explosion/spacy: v3. 7.2: Fixes for apis and requirements, Zenodo (2023).
[27] D. R. Johnson, J. C. Kaufman, B. S. Baker, J. D. Patterson, B. Barbot, A. E. Green, J. van Hell, E. Kennedy, G. F. Sullivan, C. L. Taylor, et al., Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling, Behavior research methods 55 (2023) 3726–3759.
[28] C. Ciaccio, A. Miaschi, F. Dell'Orletta, Evaluating lexical proficiency in neural language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 1267–1286. URL: https://aclanthology.org/2025.acl-long.64/. doi:10.18653/v1/2025.acl-long.64.
[29] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The annals of mathematical statistics (1947) 50–60.
[30] G. Chierchia, Reference to kinds across language, Natural language semantics 6 (1998) 339–405.

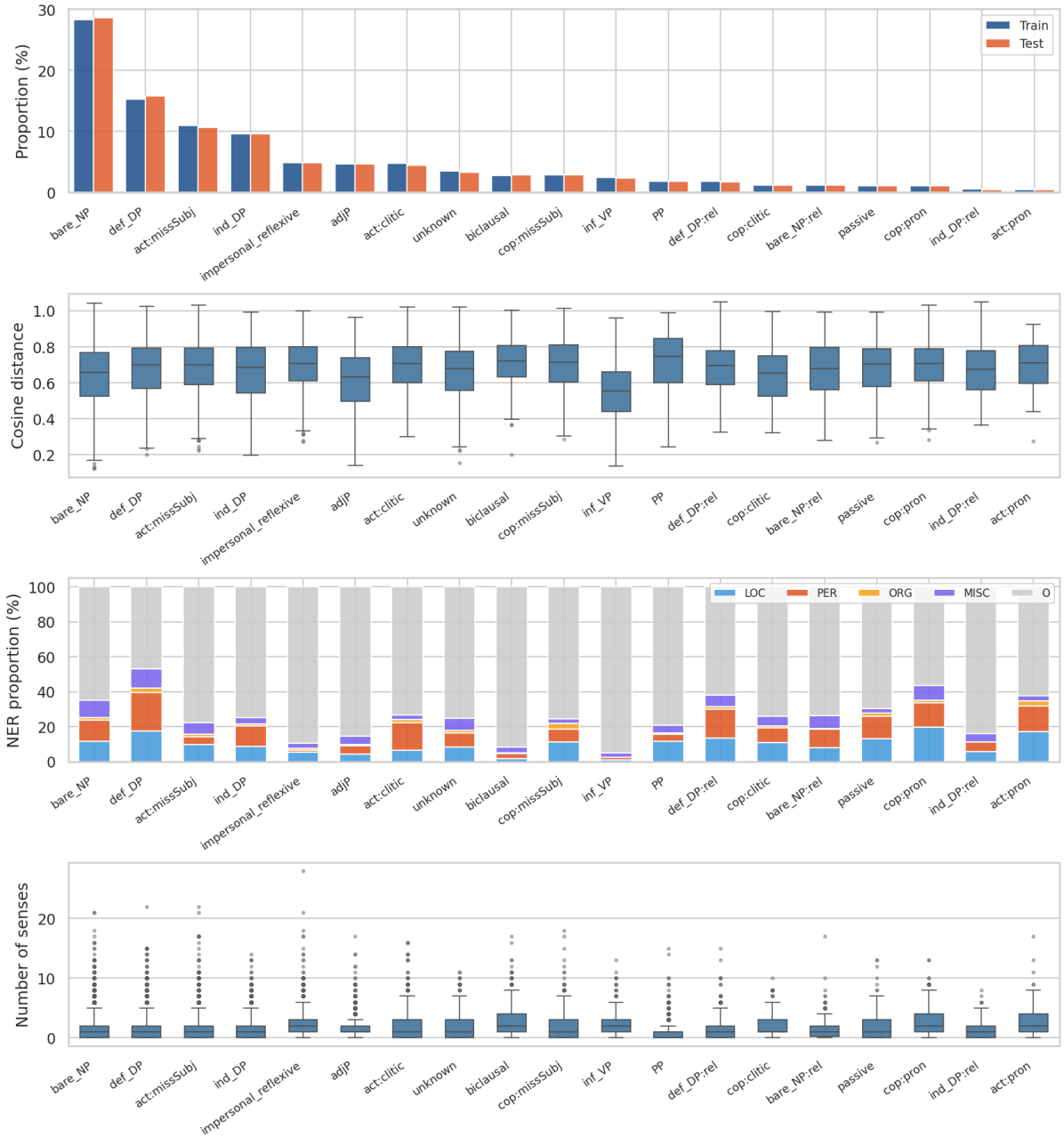

Figure 3: Distribution of syntactic categories in train (original) and test sets (a), cosine distance between clue and answer embeddings (b), proportion of NER entity types in the clue and answer (c), and number of word senses of the answer (d), broken down by syntactic category. Categories are sorted by test set frequency.

A. Linguistic Profile of the Dataset

Figure 3 reports the distribution of syntactic categories in the train and test sets of the dataset; cosine distance between clue and answer embeddings; proportion of NER entity types in the clue and answer; number of word senses of the answer. Regarding the annotation of word senses and, specifically, the stemming strategy, we report the following statistics in terms of coverage, stemming fallback and failures: for the tests set, comprising 20821 answer, 43% of them were mapped directly to the dictionary entries, 24% were mapped using the stemming fallback, 15% were too short ($\leq$ 3 chars) and 17% were unresolved.

macro_cat	syn_cat	description
nominal	bare_NP	NP without article
nominal	def_DP	DP with definite article
nominal	ind_DP	DP with indefinite article
nominal	bare_NP:rel	bare NP modified by a Relative Clause (RCl)
nominal	def_DP:rel	definite DP modified by a RCl
nominal	ind_DP:rel	indefinite DP modified by a RCl
adjectival	adjP	Adjectival Phrase (AP)
adjectival	adjP:pron	AP with a demonstrative pronoun (dem pron)
prepositional	PP	Prepositional Phrase
infinitive	inf_VP	verb phrase (VP) in infinitival form
copular	cop:missSubj	copular sentence with subject omission
copular	cop:clitic	copular sentence with predicate omission
copular	cop:pron	copular sentence with subextraction and dem pron
verb_pred	act:missSubj	sentence with active verb and subject omission
verb_pred	act:clitic	sentence with active verb and object omission
verb_pred	act:pron	sentence with active verb, with subextraction and dem pron
verb_pred	impersonal_reflexive	sentence with impersonal or reflexive verb
verb_pred	passive	sentence with verb in passive form
relativeCl_isolated	relativeCl:isolated	bare RCl (relative pronoun as root node)
biclausal	biclausal	sentence composed of two clauses

Table 3
Syntactic categories employed in the study.

	Macro-category	Category
Accuracy	0.85	0.82
Errors (n)	15	18

Table 4
Agreement between automatic and manual syntactic categorization on a stratified sample of 100 clue-answer pairs, at the macro-category and category level.

Macro-category	N. of items	Accuracy
unknown	4	0.00
adjectival	8	0.63
relativeCl_isolated	3	0.67
infinitive	4	0.75
copular	10	0.80
verb_pred	25	0.80
nominal	38	0.95
biclausal	4	1.00
prepositional	4	1.00

Table 5
Agreement between automatic and manual annotation at the macrocategory level.

B. Syntactic Categorization & Error Analysis

Categories. Table 3 reports the macro and micro syntactic categories used for annotating the dataset, along with a brief description.

Category	N. of items	Accuracy
unknown	4	0.00
adjP:pron	3	0.33
passive	3	0.33
cop:pron	3	0.33
act:missSubj	9	0.67
relativeCl:isolated	3	0.67
inf_VP	4	0.75
adjP	5	0.80
bare_NP	10	0.80
biclausal	4	1.00
cop:clitic	3	1.00
cop:missSubj	4	1.00
bare_NP:rel	3	1.00
act:pron	3	1.00
PP	4	1.00
act:clitic	5	1.00
ind_DP	8	1.00
impersonal_reflexive	5	1.00
def_DP:rel	4	1.00
def_DP	10	1.00
ind_DP:rel	3	1.00

Table 6
Agreement between automatic and manual annotation at the fine-grained category level.

Error Analysis. Imprecise categorizations and unrecognized items - included in the *unknown* category - are caused by errors in the spaCy parser. Most of them have been bypassed through some heuristic rules. To avoid the misclassification of verbs as nouns - as in *Chi l'ammazza non commette un reato* ('Whoever kills it does not commit a crime'), where *ammazza* ('kills') is incorrectly parsed as a NOUN despite being a finite verb - we specified that a noun appearing immediately after a clitic must be reclassified as a verb. Other errors stem from incorrect POS tagging: the clue *Ingannate da false promesse* ('deceived by false promises') is categorized as bare_NP instead of adjP because *ingannate* ('deceived') is parsed as a PROPN, and similarly the clue *Colpì Martin Lutero* ('It struck Martin Luther') is misclassified for the same reason. Analogous issues arise with an adjP clue like *Altezzosa* ('Haughty'), parsed as PROPN instead of ADJ, and with an act:clitic clue like *La traghetta Caronte* ('Charon ferries it'), where *traghetta* is parsed as a NOUN and categorized accordingly. These cases are particularly challenging to address with heuristic rules, since the short or absent syntactic structure leaves little context to intervene on. Further challenges concern reduced RCl which are difficult to categorize as such due to the absence of an overt relative pronoun, as in the clue *Attigui, detto dal burocrate* ('Adjacent, as said by the bureaucrat') incorrectly categorized as bare_NP:rel (*Attigui* ('Adjacent') is parsed as NOUN instead of ADJ). Finally, elliptical structures like *Il più previdente fra gli insetti* ('The most far-sighted among the insects') - where the root corresponds to an elided noun (*Il più previdente [insetto] fra gli insetti*, 'The most far-sighted [insect] among the insects') - represent a categorization challenge, as the ellipsis of the noun following the determiner makes it difficult to assign the correct syntactic category.

C. Lexico-Semantic Categorization Results

We report in Table 7 the per-system Spearman correlations between the rankings of syntactic categories by MRR@10 and their ranking by the rank-biserial effect size obtained from Mann–Whitney U tests comparing the cosine distance distribution of hits and misses at Acc@10.

Model	ρ	p-value
CrosswordSpace	-0.807	< 0.001
C.E. SBert	-0.649	0.0027
Cross++	-0.798	< 0.001
AC-DG	-0.747	0.0002
FFT-UniBa	-0.737	0.0003
MINDS	-0.861	< 0.001
UNIBA	-0.373	0.1162
Unitor	-0.633	0.0036

Table 7
Per-system Spearman correlation ρ between the ranking of syntactic categories by MRR@10 and their ranking by the rank-biserial effect size r obtained from Mann–Whitney U tests comparing the cosine distance distributions of hits and misses at Acc@10.