From Form to Meaning: Interlingua Sense-Alignment of Offensive Language with LLMs

Maria Alexandra Roussopoulou
NKUA, Greece
alexrousso@uoa.gr

Stella Markantonatou
ILSP / Athena R.C., Greece
Archimedes / Athena R.C., Greece
stiliani.markantonatou@gmail.com

Abstract

This paper presents a methodology that uses LLMs to align multilingual offensive lexicons at the sense level. Lexicons of different structures and origins in Arabic, Bulgarian, Modern Greek, French, and Italian have been aligned directly without pivoting through English. The Modern Greek lexicon is LLM-generated, and the other four lexicons are WordNet-compatible. For inter-language alignment of senses, an LLM-as-a-judge rubric was used over lemma–definition–example triples. The LLM makes 2.87M pairwise comparisons and yields 31 strict global-sense categories. The paper discusses the challenges involved in sense alignment tasks. The resource is available to support downstream applications such as Machine Translation and cross-lingual hate-speech detection.

1 Introduction

NLP research has devoted substantial effort to detecting abusive and hateful language in social media (Gevers et al., 2022; Tanev, 2024). While hate speech is supported by clearer legal or policy definitions, offensive language is harder to define because perceptions of offensiveness are subjective and culturally dependent (Kogilavani et al., 2023; Korre et al., 2025; Sariyanto et al., 2025; Loftus et al., 2025; Korre et al., 2024). This work follows Mnassri et al. (2024) who treat offensive language as an umbrella category that includes hate speech and other aggressive or derogatory expressions targeting individuals or groups.

These research directions face persistent challenges, particularly in interlingual settings. The interpretation of offensiveness is subjective and often leads to annotation disagreement (Gevers et al., 2022), while datasets and lexical resources remain unevenly distributed per language, with most resources concentrated in English and many languages being under-resourced (Korre et al., 2024; Gevers et al., 2022). As a result, multilingual approaches often rely on cross-lingual transfer methods. Prior work shows that cross-lingual embeddings and transfer learning can project offensive-language detection models trained on English to low-resource languages (Ranasinghe and Zampieri, 2020, 2021b), but these approaches depend heavily on English data and may fail to capture language-specific meanings or cultural nuances.

Another challenge arises from cultural and pragmatic variation. Offensive expressions are strongly shaped by cultural context, meaning that the same lexical form may convey different insulting meanings in different languages and communities (Mnassri et al., 2024; Korre et al., 2024; Usman et al., 2025; Dmonte et al., 2024). For example, ChatGPT 5.2 translated the following Modern Greek paragraph into English:
Πάλι το παίζει άρρωστη για να μην κάνει δουλειές. Το έχει παρακάνει το γαϊδούρι.
Pali to pezi arosti gia na min kani doulies. To exi parakani to gaidouri.

The model returned: "She's pretending to be sick again so she won't do any chores. She's really overdone it—the donkey."

This translation misses the intended insult: in Modern Greek, γαϊδούρι *gaidouri* 'donkey' is used to describe someone as rude/ill-mannered, while in English **donkey** to describe someone as stupid/stubborn. To investigate whether sense-level alignment can resolve this mismatch, the literal animal senses were aligned between them (γαϊδούρι ↔ *donkey*) and the Modern Greek metaphorical sense was aligned with the English **jerk**[1]. The English figurative sense of **donkey** 'stupid' was left unaligned. With this information, the model produced the following translation: "She's playing sick again, so she won't have to do chores. The jerk has

[1] An insulting term for someone who is rude, obnoxious, or behaves badly toward others.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 63–75

really overdone it." The model also explained that γαϊδούρι *gaidouri* 'donkey' in this context is used metaphorically as an insult referring to someone rude or obnoxious, corresponding to the aligned English sense *jerk*, rather than the literal animal donkey or the English figurative sense 'stupid'.

The example above suggests that interlingual sense alignment can guide LLMs toward translations that preserve the intended offensive meaning. Recent research (Dmonte et al., 2024) has shown that Machine Translation (MT) can significantly influence offensive language detection, and mistranslations can lead to incorrect interpretation of abusive content.

However, resources supporting sense-level interlingual alignment of offensive lexicons remain scarce. To address this gap, this work uses LLMs to align lexicons of different languages and different origins. As a case study, the lexica in the HurtNet[2] collection are used, namely Arabic, Bulgarian, French, Italian, and Modern Greek. The method uses LLMs to compare lemma–definition–example triples in their original languages and enables the alignment of meanings without semantic drift and information loss. The resulting resource supports more reliable interlingual comparison of offensive expressions and may benefit downstream applications such as safer machine translation and multilingual hate-speech detection.

A major bottleneck concerns Modern Greek, for which no reliable lexical resource provides consistent definitions and usage examples for offensive language. Existing resources, including Greek WordNet, do not adequately cover slang or offensive meanings. To overcome this limitation, a sense-level offensive lexicon for Modern Greek was created using LLMs, addressing the lack of resources for low-resource languages (Mnassri et al., 2024) while preserving meaning in the original language and avoiding English-based pivoting.

This paper is structured as follows: Section 2 presents related work. Of the languages studied, Modern Greek lacked an offensive lexicon: Section 3 describes the development and evaluation of the Modern Greek dataset. Section 4 presents and evaluates the proposed sense alignment method. The paper closes with a discussion of the results. The code and final prompts used in this work are publicly available in the project repository[3].

[2]https://github.com/valeriobasile/hurtlex/tree/hurtnet/hurtnet_data

[3]https://github.com/aroussopoulou/Interlingua-Sense-Alignment-of-Offensive-Language-with-LLMs

2 Related work

Most NLP work on offensive language and hate speech focuses on detecting such content, typically in social media corpora. Recent systematic reviews and multilingual surveys explore a wide range of modeling strategies, including Machine Learning (ML) with cross-lingual feature extraction (i.e., LASER/MUSE-style representations with logistic regression) (Ranasinghe and Zampieri, 2020; Aluru et al., 2020), deep neural architectures (e.g., CNN–LSTM and character-level CNNs, as well as cross-lingual capsule-style sequence models) (Jiang and Zubiaga, 2021; Ranasinghe and Zampieri, 2021a), Transfer Learning (TL) with multilingual transformers (e.g., mBERT/XLM-R fine-tuning and cross-lingual contextual embeddings) (Ranasinghe and Zampieri, 2020; Ghadery and Moens, 2020; Pamungkas et al., 2021), and MT/ pivot- based transfer (translate-to-pivot training or evaluation to enable cross-lingual reuse of models and labels) (Ibrohim and Budi, 2019; Pamungkas et al., 2023). This body of work has been evaluated on several widely used benchmark resources, such as OffensEval-2020 (Zampieri et al., 2020), MLMA (Ousidhoum et al., 2019), CONAN (Chung et al., 2019), and Multilingual HateCheck (Röttger et al., 2022).

Along with detection-focused work, interlinguistic sense-level alignment of offensive language is important for reliable MT and for cross-cultural analysis. The task is challenging because offensiveness is not inherent in words but it is shaped by the target and by cultural and interactional context, so the same form can be perceived differently in different communities (Culpeper, 2011). In addition, single-language pivoting (e.g., translating everything into English) can blur or disturb offense, because translation may fail to preserve socio-cultural nuance and pragmatic force (Chan et al., 2024). Inter-language sense alignment seems to be a better approach.

In general-language lexicons, sense alignment is typically approached with explicit semantic anchors: WordNet-based infrastructures, such as the Global Wordnet Grid with the Collaborative Interlingual Index (CILI), were introduced to support multilingual sense linking while preserving concepts when a single-pivot inventory is used. (Vossen et al., 2016; Bond et al., 2016). A complementary, high-coverage option is BabelNet, a

multilingual semantic network that integrates WordNet and Wikipedia and is widely used as an interlingual 'sense graph' for mapping lexicalizations across languages (Navigli and Ponzetto, 2010). FrameNet is another vocabulary that represents meaning, this time in terms of frames, their participant roles (frame elements), and annotated examples (Baker et al., 1998). Recent work has tried to align senses across languages by turning Wiktionary into a knowledge graph (DBnary in OntoLex-Lemon (Sérasset, 2015; Cimiano et al., 2016)) and then using multilingual language models to compute similarity and automatically link sense-to-sense across languages (Sérasset, 2025).

For inter-linguistic sense-level alignment of offensive terms, lexica should provide sense-separated meanings and coverage of all offensive senses of polysemous lemmas; lemmas should be in their original languages to avoid translation-induced meaning loss. In this respect, HurtLex (Bassignana et al., 2018) is a strong backbone: it is curated, multilingual, and already covers the five languages of this work (Arabic, Bulgarian, Modern Greek, French, Italian) by aligning lemmas in their original language. The main bottleneck is Modern Greek: although resources such as major dictionaries and Greek WordNet can support general sense inventories, they do not consistently provide reliable, sense-separated definitions for slang/offensive uses, and pragmatic information (e.g., target dependence) is rarely encoded (Centre for the Greek Language; Wiktionary contributors; Bond and Foster, 2013; LOD Cloud). SLANG.gr was not adopted as the main reference resource for definitions because it is user-generated and its entries are not lexicographically uniform (Xydopoulos et al., 2009). Given these limitations, the approach for Modern Greek was to generate definitions at sense-level with a strong LLM, following recent lexicography-oriented work showing that LLM outputs can approximate lexicon content of expert-style (de Schryver, 2024; Phoodai and Rikk, 2023; Jakubíček and Rundell, 2023; Fedorova et al., 2024; Han et al., 2024; Poelman and de Lhoneux, 2025; Meconi et al., 2025; Periti et al., 2024; Pham et al., 2025). In this way, a generic definition prompt (e.g., 'What does the word X mean?' / 'Generate a dictionary entry for X') can be refined into a task-specific prompt that explicitly enforces sense separation, target dependence, register, and example usage.

3 Modern Greek Dataset Construction

3.1 Pilot Study and Target Taxonomy

From the cleansed edition of HurtLex-EL (Stamou et al., 2022), 37 Greek lemmas were selected as a pilot set for a prompt producing consistent sense-specific definitions. This was necessary because several lemmas are polysemous (i.e., they have multiple senses). The following Greek lexical sources were used in this phase: Babiniotis' dictionary (Babiniotis, 2024), the Portal for the Greek Language (Triantafyllidis' Dictionary) (Centre for the Greek Language), and Wiktionary (Wiktionary contributors).

The definitions took into account the relationship between the interpretation of an insult and its target, as changing the target often changes the offensive meaning (Stamou et al., 2022; Bolinger, 2015; Camp, 2013). Next, targets were grouped into persistent entities (Person, Object) and temporal entities (State, Repetitive Action, Event, Behavioral Trait) (CIDOC CRM Special Interest Group, 2021). Each lemma sense was anchored to one (or more) of these six target classes if senses were available for them. As an example, we provide the Greek word βρώμα *vroma* 'filth/ stink', which functions as an insult for an **unethical person** if the target is a person, refers to **an immoral act or a scandal** when it characterizes a behavior, to **uncleanliness or dirtiness** when it targets a place or a situation, and to **strong unpleasant smell** when it targets an object.

3.2 Prompt Design

Prompt engineering was performed using GPT-o3, selected for its strong reasoning and multi-turn instruction (OpenAI, 2025), as well as its cost profile (OpenAI, n.d.). Previous work shows that GPT-based models can produce dictionary-style definitions that are comparable to human lexicography and are not simple reproductions of existing dictionary text (Pham et al., 2025). The final prompt was written in Modern Greek and was developed with 15 prompt iterations in the ChatGPT environment with the 37 pilot lemmas. It enforced: (i) **sense listing** per lemma, (ii) an one-sentence **genus differentia** definition per sense that avoids the headword and its derivational family (Atkins and Rundell, 2008), uses common vocabulary and notes the typical context or tone in which each offensive word is used, (iii) **three usage examples** per sense following GDEX criteria (typicality, informativeness,

readability) (Kilgarriff et al., 2008) with diversity of settings, and (iv) assignment of each sense to a **target class**, guided by explicit decision rules (Wei et al., 2022). The prompt also specified a fixed output schema, which returns the lemma, part of speech (POS), sense with the target class, definition, synonyms, and examples, repeating the same format for each additional sense.

3.3 Lexicon Generation

The prompt was used to generate the definitions in a zero-shot setting via the OpenAI API using GPT-o3 by processing the 737 HurtLex-EL lemmas in batches of 20 per request. The produced 802 draft definitions were manually post-edited; 536 edits were logged (8 incorrect senses, 307 missing/merged senses, and 221 minor edits affecting definitions, categories, synonyms, or examples), resulting in 1,109 final, sense-level definitions. Manual intervention was specifically applied to cases where a lemma targeted distinct protected groups or carried different levels of offensive intensity.

For most definitions, an authentic Greek example matching the intended sense was selected from the AIKIA corpus (Markantonatou et al., 2024) of Modern Greek offensive language. AIKIA did not supply suitable matching examples for approximately 200 definitions. To address this gap, examples were collected from Twitter/X and assigned offense scores by four annotators using the Best–Worst Scaling method (Basile and Cagnazzo, 2021), which is the offense score assignment method also used with the AIKIA dataset.

3.4 Evaluation

Two complementary evaluations were carried out: (1) An expert review by three lexicographers using a structured questionnaire covering 32 senses assigned to 20 lemmas, following a multi-level evaluation grid for AI-generated dictionary entries (Evert et al., 2024). Experts rated six qualitative criteria on a 1–5 scale (1 = poor, 5 = excellent): target class, clarity of meaning, plain language without headword/morphological family, linguistic correctness, example quality, and synonym choice. In addition, one quantitative criterion checked sense coverage using a binary 1–2 scale (1 = missing, 2 = covered). Scores were aggregated to 2,388/2,920 (81%), with 36/40 (90%) coverage. The expert evaluation highlighted four systematic issues: abandoned older forms of the language were presented as currently used, English-pattern interference, leakage of evaluative tone, and sense drift in the generated examples.
(2) Three non-expert lexicographers assessed lexicon coverage on ParlaMint-GR (Greek parliamentary proceedings) (Centre for the Greek Language). Offensive sentences were first detected automatically with AIKIA (Markantonatou et al., 2024), which assigns continuous offensiveness scores, and offensive words in those sentences were then annotated manually. The analysis identified 1,229 offensive word tokens, of which 125 were attested in the Modern Greek lexicon (10.2% coverage); among the attested lemmas, three corpus-attested senses were missing from the lexicon. This relatively low coverage is expected, since parliamentary discourse is a highly formal and institution-specific register in which offensiveness is often expressed indirectly and differs from everyday insult vocabulary.

3.5 Final Dataset

The final dataset for sense alignment was assembled from the HurtNet branch of the public repository and contains 2,784 sense entries from five languages, namely Arabic, Bulgarian, Modern Greek, French, Italian. Only the shared alignment fields were used: ID (unique lemma per language), Language, POS, lemma (surface form), Definition ID (unique sense entry), Definition (sense description), and Example. For analysis only, lemmas and definitions were machine-translated into English to facilitate cross-lingual comparison.

4 The alignment methodology

Because the five languages contain unequal numbers of sense entries (Arabic: 438, Bulgarian: 637, Modern Greek: 1109, French: 330, Italian: 271), cross-lingual alignment must rely on large-scale pairwise comparisons to avoid missing rare but valid matches. In this setting, LLM-as-a-Judge offers a practical way to compare senses in their original languages, enabling millions of pairwise comparisons and producing concise rationales that can be parsed automatically and spot-checked by humans (Zheng et al., 2023). However, previous works have shown that LLM-based judges can be biased or unstable, because of position and verbosity effects, prompt sensitivity, and nondeterminism (Zheng et al., 2023; Shi et al., 2025; Ye et al., 2025; Stureborg et al., 2024).

4.1 Judging Prompt

Taking into account these constraints, the judging prompt was designed in English as a rubric-based decision task A. The model was instructed to act as an expert judge in offensive language and multilingual lexicography and to decide whether two terms express the same offensive sense, based on their meaning and use. The focus on **meaning and use** is important because abusive-language research shows that labels can become inconsistent when the phenomenon is not explicitly specified with clear decision criteria (Vidgen and Derczynski, 2020; Sulpizio et al., 2024). Therefore, each sense was provided with its lemma, definition, and usage example in the original language, to avoid translation-mediated shifts that may dilute culturally specific offensiveness (Chan et al., 2024). For each pairwise comparison, the model had to assign one label out of the following three ones: **merge** (same sense), **related** (clear overlap but not identical), or **unrelated** (no meaningful overlap). The label merge was used when the two entries matched in three aspects: **function** (same communicative role, such as a direct insult vs. a curse), **target** (aimed at the same type of person/group/situation), and **core insult with similar severity** (the same main derogatory or roughly equally harsh insult) (Määttä, 2023). Finally, the model was required to follow a fixed output format. For each pair, it returned one structured line with the decision (*merge*/*related*/*unrelated*), a confidence score, a label (when applicable), and a short rationale explaining the choice, so the results could be saved directly as JSONL, parsed reliably, and quickly spot-checked.

4.2 Alignment

Each entry from the 2,785 entries of the multilingual lexicon was treated as a sense node identified by its language and its definition ID:

$$v = (\ell, r) \tag{1}$$

where $\ell \in \{\texttt{ar}, \texttt{bg}, \texttt{el}, \texttt{fr}, \texttt{it}\}$ and r is the `Definition_ID` in the source file. For example, $v = (\texttt{el}, 1320)$ denotes the Greek sense with the Definition ID 1320. After representing the lexicon entries as sense nodes, we generated pairwise comparison tasks for all 10 language pairs by comparing every sense in ℓ_a against every sense in ℓ_b for each (ℓ_a, ℓ_b), i.e., using a Cartesian product to ensure complete coverage despite unequal numbers of senses across languages. Each comparison was written as one JSON object per line in a batch JSONL file. As a result, each request contained an alignment ID and a body with the chat-completion fields used in this work (model, temperature, response format, and the system/user messages) (Mistral AI, 2026b). Requests were split into batch jobs of 10,000 comparisons (Mistral AI, 2026a) and executed with mistral-small-latest (Mistral AI, 2025) and temperature=0 to keep the judge's decisions as stable and reproducible as possible across millions of pairwise comparisons. The output files were downloaded and merged into a single JSONL file for each language pair.

4.3 Alignment' Statistics

The LLM's pairwise decisions are summarized in Table 1. Across all 2,872,327 comparisons, the model labeled 0.75% as merge, 19.0% as related, and 80.2% as unrelated. These results were expected given that strict one-to-one equivalences across languages are rare in offensive vocabulary. The hardest language pair to align strictly was Bulgarian–Greek (MERGE 0.3%), while the highest MERGE rates were observed for Arabic–Italian (2.2%) and Arabic–French (1.8%).

4.4 Global Senses

At this stage **global senses**, i.e., meaning groups shared across all five languages were derived from the pairwise LLM decisions stored in the 10 language pair JSONL files. This was necessary because pairwise agreement is not transitive: it is possible for A to merge with B and B to merge with C, but A and C are only *related* (or even *unrelated*). This issue has been observed in work on LLM-as-a-judge and pairwise comparisons, which notes that such methods rely on transitivity and reports that LLM judges can exhibit non-transitive preferences (Xu et al., 2025). For this reason, global senses were extracted using a **strict unification rule**: a 5-language tuple $(v_{ar}, v_{bg}, v_{el}, v_{fr}, v_{it})$ was accepted only if **all 10 internal language pairs** inside the tuple were labeled as *merge*. Since enumerating all possible cross-language combinations was infeasible, candidates were generated using a pivot language p. Within each pivot-language run, the algorithm goes through every **pivot sense node** v_p in that language and looks up which senses in the other languages have a *merge* link with it. For any language ℓ, the **merge-neighborhood** of v_p is

Language pair	Merge (%)	Related (%)	Unrelated (%)	Total
ar–bg	1.1	22.4	76.5	278,919
ar–el	0.9	19.8	79.4	485,657
ar–fr	1.8	30.5	67.6	144,511
ar–it	2.2	28.7	69.1	118,679
bg–el	0.3	15.7	84.0	706,238
bg–fr	0.6	26.3	73.1	210,131
bg–it	0.6	27.4	72.0	172,508
el–fr	0.6	12.8	86.6	365,879
el–it	0.5	13.3	86.1	300,396
fr–it	0.6	10.3	89.1	89,409
All pairs	0.75	19.0	80.2	2,872,327

Table 1: Pairwise alignment decisions by language pair (percentages).

defined as:

$$N_\ell(v_p) = \{\, v_\ell \in V \mid (v_p, v_\ell) \in E \,\}. \quad (2)$$

Here, V is the set of all sense nodes and E is the set of all merge links predicted by LLM. Using these **merge-neighborhood**, the algorithm builds candidate 5-tuples by picking one neighbor from each of the four non-pivot languages and combining them with the pivot sense. Each candidate is then verified by the **strict unification rule**. This process is repeated once per pivot language, where S_p denotes the set of **verified strict tuples** obtained with language p as the pivot. The pivot runs produced **294,835** in total, with $S_{ar} = 23,621$, $S_{bg} = 82,514$, $S_{el} = 72,882$, $S_{fr} = 55,622$ and $S_{it} = 60,196$. All pivot outputs are then combined and deduplicated so that the same tuple appears only once, resulting in **191,083 unique strict tuples**. To make the **unique strict tuples** easier to interpret as global senses, we grouped them using the **4/5 similarity rule**. Two strict 5-tuples are placed in the same category when they match in four out of five languages and differ in only one language, which typically indicates an alternative lexical choice in that language rather than a different meaning. For example, the tuples $T = (ar : 34, bg : 930, el : 1323, fr : 2222, it : 2674)$ and $T' = (ar : 34, bg : 930, el : 1323, fr : 2222, it : 2783)$ match in Arabic, Bulgarian, Greek, and French, and differ only in Italian; under the 4/5 similarity rule, they are grouped into the same global sense, and the two Italian entries are treated as alternative lexical choices within that category. Grouping is applied transitively, meaning that if T matches T' in 4/5 languages and T' matches T'' in 4/5 languages, then all three are assigned to the same cluster. Overall, the grouping procedure produced 31 global sense categories[4]. Finally, to produce a reliable and compact structured set of global senses, each category was organized first by target (person, behavior/stance, or state) and then by part of speech.

4.5 Evaluation

To evaluate the quality of the sense-alignment method, two complementary evaluations were conducted: (1) a comparison against the HurtLex multilingual resource alignment as a reference baseline (Bassignana et al., 2018), and (2) a bilingual human judgment study to directly assess whether the LLM's merge decisions satisfy the three criteria encoded in the judging prompt (**function**, **target**, and **core insult with similar severity**).

(1) For a fair comparison, the publicly released HurtLex alignment table was restricted to the five HurtNet languages (Arabic, Bulgarian, Greek, French, Italian) and further filtered to alignment IDs with 5/5 coverage, i.e., IDs whose lemmas appear in the HurtNet lexicon in all five languages. Under this filtered setting, HurtLex yields 52 lemma-level alignment categories, whereas HurtNet produces 31 sense-based global-sense categories: HurtLex aligns lemmas, while HurtNet aligns sense entries using lemma, definition, and example, and further organizes results by target and part of speech. As a result, HurtLex is more exposed to false matches due to polysemy and missing context, while the LLM-based sense alignment can lead to missed merges (near-misses) or occasional false merges when decisions are sensitive to prompt framing or the wording of definitions/examples (Zheng et al., 2023; Shi et al., 2025). Table 2 reports on the coverage and over-

[4]The resource created with alignment is available at https://osf.io/4sahk/overview?view_only=06b5bb13043c498c97b7af63244f013d

Language	HurtLex aligned	HurtNet aligned	Common	HurtLex cov.	HurtNet cov.
Arabic	98	138	50	22.4%	31.5%
Bulgarian	107	112	47	18.0%	18.9%
Greek	125	113	48	17.5%	15.8%
French	83	68	38	37.6%	30.8%
Italian	83	65	40	34.3%	26.9%
Total	496	496	223 (45%)	22.5%	22.5%

Table 2: Coverage and overlap between HurtLex and HurtNet (filtered to the five HurtNet languages).

Language pair	Yes	No	Total	Accuracy (%)
AR–IT	15	5	20	75.0
EL–IT	20	0	20	100.0
EL–FR	16	4	20	80.0
EL–BG	17	3	20	85.0

Table 3: Bilingual-speaker evaluation of 20 LLM-proposed sense alignments per language pair (Yes = same sense; No = different sense).

lap between the filtered HurtLex and the HurtNet alignment. In Arabic and Bulgarian, HurtNet aligns more lemmas than HurtLex (Arabic: 138 vs. 98; Bulgarian: 112 vs. 107), indicating that the sense-based pipeline can recover additional matches once definition and usage context are taken into account. In French, Italian, and Modern Greek, HurtNet aligns fewer lemmas than HurtLex. This result does not necessarily indicate lower coverage; rather, it reflects stricter sense-level matching that avoids lemma-level merges when meanings diverge across languages. Overall, the two methods yield the same total number of aligned lemmas in this filtered setting (496), of which 223 lemmas (about 45%) are shared. This suggests that HurtLex is a useful starting point for cross-lingual alignment, but that sense-aware alignment can refine or correct matches when polysemy and cross-lingual meaning shifts are present.

(2) In the second evaluation, the goal was to assess whether the LLM-as-a-judge classified sense alignments correctly according to the judging-prompt criteria. Four bilingual speakers were recruited, one for each of the available language pairs. Each participant completed a short questionnaire with 20 aligned sense pairs. For every pair, the lemmas and their definitions in the original language were shown, and the annotator judged whether the two senses were the same (Yes) or different (No). Language pairs of Arabic–Italian (AR–IT), Modern Greek–Italian (EL–IT), Modern Greek–French (EL–FR), and Modern Greek–Bulgarian (EL–BG) were evaluated; when a pair was marked as No, annotators provided a brief justification. As shown in Table 3, EL–IT achieved perfect agreement (20/20, 100%), while the remaining pairs exhibited 3–5 errors (accuracy 75–85%), with AR–IT obtaining the lowest score (15/20, 75%). A recurring source of disagreement was the borderline case of a core insult with comparable severity but not identical meaning. For instance, Arabic بخيل *bakhiil* 'stingy' was incorrectly merged with Italian **povero** 'poor', which denotes economic hardship rather than stinginess. Similarly, the Greek word παλιάνθρωπος *palianthropos* 'vile person' was merged with the French word **salaud** 'bastard', a stronger and more vulgar insult. Lastly, annotators flagged a lemma-choice issue in French: although the entry uses **porc**, whose literal interpretation is 'pork' (meat), the more idiomatic offensive term for the intended insulting sense is **cochon**; the definition nonetheless correctly captures the insulting meaning.

5 Results & Analysis

The sense-based alignment yields 31 global-sense categories, each intended to represent a shared insulting meaning across languages. Insults for low intelligence/stupidity, comprise **195** sense entries across the five languages followed by insults for deception/lying (**103**), worthlessness/being trash (**48**), maliciousness (**41**), and sex work (female-directed slurs) (**34**). Beyond these high-frequency clusters, many categories are smaller and more specific (often 1–3 entries per language), such as insults invoking animals (e.g., 'dog' used to denote a worthless or contemptible person) or ideology-

based insults (e.g., 'fascist'), as well as a distinct cluster for mess/chaos that is realized via both general terms (e.g., chaos) and colloquial extensions (e.g., French **bordel** 'mess/ brothel (pej.)' and Greek μπουρδέλο *bourdelo* 'mess/ brothel (pej.)').

To identify failure points in strict multilingual merging, we inspected cases in which a fully merged **4-language subtuple** exists but the fifth language is missing. The procedure selects instances with exactly four languages present, and all the six internal language pairs within that 4-language subtuple are labeled *merge*, and then records which language is absent most often. The resulting distribution is clearly uneven: Italian is the most frequently missing language (54,148 cases), followed by Modern Greek (50,904), and Bulgarian (47,559). French and Arabic are missing in fewer cases (34,612 and 25,852, respectively). Regarding Italian, the high missing language rate likely reflects lexicalization gaps (Khishigsuren et al., 2022; Li et al., 2024), where the Italian inventory lacks a direct lemma for a meaning present in the other languages, so it cannot complete an otherwise consistent 5-language merge group. In contrast, Modern Greek is frequently missing despite having the largest sense inventory, which suggests a sense-granularity mismatch: Greek meanings may be split more finely and described with narrower definitions, making strict merge decisions harder and increasing cases that are labeled only as related (Bevilacqua et al., 2021a,b). Overall, merge participation is high across languages ($\approx 90\%$ of senses merge at least once; Arabic 97.7%, Bulgarian 93.2%, French 92.4%, Italian 91.8%, Modern Greek 84.5%), leaving a small but informative set of never-merge senses that are examined next as potentially language-specific or hard-to-align meanings. For example, French **trou** is used as familiar slang for 'prison' (e.g., au trou), a strongly register-bound meaning that may not map cleanly onto a single strict counterpart in other languages, while Bulgarian шваба *shvaba* is a colloquial (often pejorative) term for a 'German', illustrating ethnonym-based offensiveness that is highly usage-dependent. In Modern Greek, δεκανέας *dekaneas* used pejoratively as 'petty authoritarian / small-power bully' and αμερικανάκι *amerikanaki* as 'naive imitator of American culture' encode culturally specific social stereotypes that are not equally lexicalized elsewhere. Similarly, Italian items tied to socially marked categorizations, e.g., the word **carne** 'meat' (usage connected to light skin) and Arabic مشرك *mushrik* 'polytheist' as a doctrinal accusation in Islam, reflect meaning components that are strongly culture and discourse-bound, which can hinder strict cross-lingual merging.

6 Conclusions

This work explores the challenges involved in developing a multilingual offensive language lexicon by aligning explicitly defined offensive senses. To this end, it introduces a methodology for sense-level alignment of offensive lexicons of Arabic, Bulgarian, Modern Greek, French, and Italian. It compares lemma–definition–example triples in their original languages using an LLM-as-a-judge rubric, enabling the identification of shared insulting meanings without relying on pivot translation. The alignment results reveal insulting senses shared by all languages as well as language-specific ones. Lexicalization gaps, differences in sense granularity, and partially overlapping meanings across languages were shown to be important impediments for the sense-alignment task.

Developing lexica of under-resourced languages is a challenge which this work addressed by developing a lexicon for the under-resourced Modern Greek. An offensive lexicon for Modern Greek was generated using LLMs, containing definitions and examples that enabled its inclusion in the interlingual alignment pipeline and helped reduce resource gaps for low-resource languages. The evaluation revealed several limitations of LLM-generated lexicographic content, including the use of outdated language forms, interference from English patterns, leakage of evaluative tone in definitions, and semantic drift in the generated examples.

Future work will primarily explore downstream applications in MT, as previous studies show that mistranslations can significantly affect offensive language detection and interpretation (Dmonte et al., 2024). Using the aligned lexicon may help reduce semantic drift and preserve the intended abusive meaning across languages. Beyond MT, the resource could also support other multilingual NLP tasks, such as multilingual hate-speech detection, cross-lingual transfer learning, and embedding-based semantic modeling. Additionally, it may serve as a pedagogical tool for second-language learning (L2) by providing accurate cross-lingual correspondences of offensive expressions and their contextual meanings.

Limitations

An important limitation is the evaluation coverage; bilingual human judgments were obtained only for a subset of language pairs because it was difficult to recruit evaluators with the required bilingual competence, so the alignment quality is not validated uniformly across all pairs. In addition, cross-language comparability is constrained by resource heterogeneity: each language relies on different lexical sources and lexicographic conventions (coverage, sense granularity, and example selection), which can influence the resulting lexicon and the alignment statistics.

Acknowledgments

We would like to express our sincere gratitude to the HurtNet team, and in particular to Valerio Basile, Adel Mahmoud Wizani and Petya Osenova, for their invaluable support.

We are also grateful to all participants who contributed to the evaluation process. In particular, we thank Mavina Pantazara and Panagiotis Krimpas for their expertise in the dictionary evaluation. We also acknowledge the contribution of the alignment evaluators, Sofia Roussopoulou, Mr Thanasis, and Nese Patrizio, as well as Eleftherios Leonidas Canterakis, Konstantinos Diamantopoulos, Aggeliki Kourou, and Maria Poulou, whose work was essential to the development of the Modern Greek dictionary dataset and the evaluation of usage examples.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *Preprint*, arXiv:2004.06465.

B. T. S. Atkins and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

George Babiniotis. 2024. *Dictionary of Modern Greek*, 6th edition. Lexicology Centre. Accessed: 2026-02-25.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Valerio Basile and Christian Cagnazzo. 2021. Litescale: A lightweight tool for best-worst scaling annotation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 121–127, Held Online. INCOMA Ltd.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 52–57, Turin, Italy. CEUR Workshop Proceedings.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021a. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021b. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pages 4330–4338.

Renee Bolinger. 2015. The pragmatics of slurs. *Noûs*, 51.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

Elisabeth Camp. 2013. Slurring perspectives. *Analytic Philosophy*, 54(3):330–349.

Centre for the Greek Language. Portal for the greek language. Website. Accessed: 2026-02-25.

Fai Leui Chan, Duke Nguyen, and Aditya Joshi. 2024. "is hate lost in translation?": Evaluation of multilingual LGBTQIA+ hate speech detection. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 146–152, Canberra, Australia. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

CIDOC CRM Special Interest Group. 2021. *CIDOC Conceptual Reference Model (CRM), Version 7.2*. ISO 21127:2023 correspondence document.

Philipp Cimiano, John P. McCrae, Paul Buitelaar, and 1 others. 2016. Lexicon model for ontologies: Ontolex-lemon. W3c community group report, W3C Ontology Lexicon Community Group. Published 10 May 2016.

Jonathan Culpeper. 2011. *Impoliteness metadiscourse*, page 71–112. Studies in Interactional Sociolinguistics. Cambridge University Press.

Gilles-Maurice de Schryver. 2024. The road towards fine-tuned LLMs for lexicography. In *Book of Abstracts of the Workshop "Large Language Models and Lexicography" (EURALEX 2024 workshop)*, pages 6–11, Ljubljana. ELEXIS Association.

Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri. 2024. On the effects of machine translation on offensive language detection. *Social Network Analysis and Mining*, 14(1):242.

Stephanie Evert, Christine Ganslmayer, and Christian Rink. 2024. Multi-level analysis as a systematic approach to evaluating the quality of ai-generated dictionary entries. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pages 317–335, Cavtat. Institut za hrvatski jezik.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Ine Gevers, Ilia Markov, and Walter Daelemans. 2022. Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands Journal*, 12:33–48.

Erfan Ghadery and Marie-Francine Moens. 2020. LIIR at SemEval-2020 task 12: A cross-lingual augmentation approach for multilingual offensive language identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2073–2079, Barcelona (online). International Committee for Computational Linguistics.

Yi Han, Ryohei Sasano, and Koichi Takeda. 2024. Definition generation for automatically induced semantic frame. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11112–11118, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Translated vs non-translated method for multilingual hate speech identification in twitter. *International Journal of Advanced Science, Engineering and Information Technology*, 9(4):1116–1123.

Miloš Jakubíček and Michael Rundell. 2023. The end of lexicography? can ChatGPT outperform current tools for post-editing lexicography? In *Proceedings of eLex 2023*.

Aiqi Jiang and Arkaitz Zubiaga. 2021. Cross-lingual capsule network for hate speech detection in social media. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 217–223, New York, NY, USA. Association for Computing Machinery.

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2798–2807, Marseille, France. European Language Resources Association.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Barcelona, Spain.

S.V. Kogilavani, S. Malliga, K.R. Jaiabinaya, M. Malini, and M. Manisha Kokila. 2023. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 81:630–633. International Virtual Conference on Sustainable Materials (IVCSM-2k20).

Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. The challenges of creating a parallel multilingual hate speech corpus: An exploration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia. ELRA and ICCL.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. *Preprint*, arXiv:2411.07417.

Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2024. Translation-based lexicalization generation and lexical gap detection: Application to kinship terms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6891–6900, Bangkok, Thailand. Association for Computational Linguistics.

LOD Cloud. Open multilingual wordnet (greek) dataset (omwn-ell). Dataset metadata page. Accessed: 2026-03-02.

Sebastian Loftus, Adrian Mülthaler, Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2025. Using LLMs and preference optimization for agreement-aware HateWiC classification. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*,

pages 538–547, Vienna, Austria. Association for Computational Linguistics.

{Simo K.} Määttä. 2023. Linguistic and discursive properties of hate speech and speech facilitating the expression of hatred: Evidence from finnish and french online discussion boards. *Internet Pragmatics*, 6(2):156–172.

Stella Markantonatou, Vivian Stamou, Christina Christodoulou, Georgia Apostolopoulou, Antonis Balas, and George Ioannakis. 2024. Aikia corpus. OSF repository.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.

Mistral AI. 2025. Mistral small 3.2 (v25.06) | mistral docs. https://docs.mistral.ai/models/mistral-small-3-2-25-06. Published: 2025-06-20. Accessed: 2026-02-26.

Mistral AI. 2026a. Batch inference | mistral docs. https://docs.mistral.ai/capabilities/batch/. Accessed: 2026-02-26.

Mistral AI. 2026b. Usage (chat completions) | mistral docs. https://docs.mistral.ai/capabilities/completion/usage. Accessed: 2026-02-26.

Khaled Mnassri, Reza Farahbakhsh, Rasoul Chalehchaleh, Pramudith Rajapaksha, Amir Reza Jafari, Guoliang Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

OpenAI. 2025. Introducing o3 and o4-mini. OpenAI Blog.

OpenAI. n.d. Pricing | openai api. https://developers.openai.com/api/docs/pricing. Accessed: 2026-02-25.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Inf. Process. Manag.*, 58(4):102544.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27:17–43.

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically generated definitions and their utility for modeling word meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Bach Pham, JuiHsuan Wong, Samuel Kim, Yunting Yin, and Steven Skiena. 2025. Word definitions from large language models. In *2025 19th International Conference on Semantic Computing (ICSC)*, page 158–162. IEEE.

Chayanon Phoodai and Richárd Rikk. 2023. Exploring the capabilities of ChatGPT for lexicographical purposes: A comparison with Oxford Advanced Learner's Dictionary within the microstructural framework. In *Proceedings of eLex 2023*.

Wessel Poelman and Miryam de Lhoneux. 2025. The roles of English in evaluating multilingual language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498, Tallinn, Estonia. University of Tartu Library.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021a. An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, 12(8). : © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

Tharindu Ranasinghe and Marcos Zampieri. 2021b. Multilingual offensive language identification for low-resource languages. *Preprint*, arXiv:2105.05996.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, and Marc Ebner. 2025. Towards explainable hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12883–12893, Vienna, Austria. Association for Computational Linguistics.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon based rdf multilingual lexical resource. *Semantic Web*. Dataset Description paper (Semantic Web Journal).

Gilles Sérasset. 2025. Towards sense to sense linking across DBnary languages. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 318–327, Naples, Italy. Unior Press.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Simone Sulpizio, Fritz Günther, Linda Badan, Benjamin Basclain, Marc Brysbaert, Yuen Lai Chan, and 1 others. 2024. Taboo language across the globe: A multilab study. *Behavior Research Methods*, 56:3794–3813.

Hristo Tanev. 2024. JRC at ClimateActivism 2024: Lexicon-based detection of hate speech. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 85–88, St. Julians, Malta. Association for Computational Linguistics.

Muhammad Usman, Muhammad Ahmad, M. Shahiki Tash, Irina Gelbukh, Rolando Quintero Tellez, and Grigori Sidorov. 2025. Multilingual hate speech detection in social media using translation-based approaches with large language models. *Preprint*, arXiv:2506.08147.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual GlobalWordnet grid. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 424–431, Bucharest, Romania. Global Wordnet Association.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wiktionary contributors. Wiktionary: The free dictionary. Website. Accessed: 2026-02-25.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating non-transitivity in llm-as-a-judge. *Preprint*, arXiv:2502.14074.

George J. Xydopoulos, Anna Iordanidou, and Anastasia Efthymiou. 2009. Recent advances in the documentation of greek slang: The case of `www.slang.gr`. In *Proceedings of the 9th International Conference on Greek Linguistics (ICGL9)*, Chicago, USA. Paper first presented at ICGL9 (October 2009); PDF distributed via OSU ICGL proceedings.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. ArXiv:2306.05685v4 (NeurIPS 2023 Datasets and Benchmarks Track).

A Judging Prompt

You are an expert in offensive language and multilingual lexicography. Your task: given TWO OFFENSIVE TERMS (Term A and Term B), decide how similar they are in sense of their offensive meaning and use.
You must choose exactly ONE of:

- **merge** → same offensive sense (one Sense)

- **related** → clearly related / near-synonyms, but not exactly the same sense
- **unrelated** → no useful offensive relation

Guidelines (short):

- Use **merge** only if they share:
  1. same offensive purpose (e.g., expletive for a bad event, insult for the same type of person),
  2. same typical target (situation/person/group),
  3. same core offensive idea and similar strength.
- Use **related** if they are in the same offensive area but differ in purpose, target, or strength.
- Use **unrelated** if their offensive domain or purpose is different, or if any link is too vague.

Be conservative:

- If between **merge/related** → choose **related**.
- If between **related/unrelated** → choose **unrelated**.

OUTPUT FORMAT (VERY IMPORTANT):
Answer with EXACTLY ONE LINE, with 5 fields separated by `|||`, in this order:
`decision|||confidence|||globalsense_id|||globalsense_label_en|||rationale_en`
Where:

- **decision**: one of `merge`, `related`, `unrelated` (lowercase)
- **confidence**: a number between 0 and 1 with at most 2 decimals (e.g., `0.87`)
- **sense_label_en**:
  - if decision is `merge` or `related`: short English label (max 12 words)
  - if decision is `unrelated`: write `null`
- **rationale_en**: 1–2 short sentences in English explaining your decision (mention purpose, target, and core offensive idea; do NOT print any slurs)

Your entire reply MUST be this single line. Do NOT add explanations before or after. Do NOT use markdown or code fences.