Are `Crescia` and `Piadina` the Same? Towards Identifying Synonymy or Non-Synonymy between Italian Words to Enable Crowdsourcing from Language Learners

Lavinia Aparaschivei[1,2], Lionel Nicolas[2], Alberto Barrón-Cedeño[1]

[1] DIT–Università di Bologna, Forlì, Italy

[2] Institute for Applied Linguistics, Eurac Research, Bolzano, Italy

{lavinianicoleta.aparaschivei, lionel.nicolas}@eurac.edu,
a.barron@unibo.it

Abstract

We introduce a method to generate candidate pairs of related Italian words sharing (or not) synonymous relations from the ConceptNet knowledgebase. The pairs are intended to generate questions for a vocabulary trainer which combines exercises to enhance vocabulary skills with the implicit crowdsourcing of linguistic knowledge about the semantic relations between words. Our method relies on the idea that pairs of synonyms in a language tend to translate to pairs of synonyms in other languages. We generated $85k$ candidate pairs of Italian synonyms that can be used to produce questions for both teaching ($3.8k$ pairs) and crowdsourcing purposes ($80k$ pairs). Follow-up efforts are however needed in order to generate a complementary set of questions.

1 Introduction

Our efforts target the automatic generation of semantically-related candidate pairs of Italian words with a focus on synonymy. We address a cold start issue for a vocabulary trainer combining exercises to enhance vocabulary skills with the implicit crowdsourcing of linguistic knowledge about the semantic relations between words.

While targeting a specific use case, our method contributes to a larger effort aimed at narrowing gaps on two fronts. On the NLP front, over the past few decades varied efforts have targeted the efficient creation, extension, and maintenance of

resources, including crowdsourcing through platforms such as Amazon Mechanical Turk (Ganbold et al., 2018; Potthast et al., 2018). Still, the subject remains an open issue. On the computer-assisted language learning (CALL) front, the automatic generation of exercise content from NLP resources is almost non-existent, despite the fact that some of these datasets encode the knowledge that learners are often tested on (e.g., lexical knowledge). This absence is probably due to differences in expectations with respect to linguistic accuracy: learning materials are usually close to perfect, whereas NLP resources rarely are. Generating content from imperfect datasets poses a challenge in terms of its suitability for learning.

We contribute to narrowing these gaps by producing data to tackle a cold start issue for a vocabulary trainer designed to both teach language and crowdsource linguistic knowledge from learners. We generate a collection of candidate pairs of Italian words tied to confidence scores, allowing to decide which pairs should be used for learning or for crowdsourcing purposes. Our method projects synonymy information in ConceptNet (Speer et al., 2017) from non-Italian onto Italian words. The obtained results show that we adequately tackle part of the cold start issue, while follow-up efforts are needed to address the remaining part.

The rest of the paper is organised as follows: Section 2 discusses the specific purpose of our method. Section 3 summarises related work. Section 4 and Section 5 describe how the candidate pairs are generated and scored. Finally, Section 6 discusses how suitable the pairs are for our specific use case and Section 7 provides closing remarks.

2 Background

Ours and previous related work (Lyding et al., 2019; Rodosthenous et al., 2019; Rodosthenous et al., 2020; Nicolas et al., 2021) all contribute to

a wider effort to research an implicit crowdsourcing paradigm built upon the idea that the curation of NLP resources and language learning are sibling endeavors (Nicolas et al., 2020). On the one hand, (NLP) researchers try to create models to "teach" a computer to process and/or produce language utterances. On the other hand, learners create a model, in the form of personal knowledge, to process and/or produce language utterances too. Allowing learners to express their knowledge can contribute to enhance an NLP resource, under specific conditions. This paradigm substitutes the expert manpower typically required to curate NLP resources with a non-expert crowd of learners.

Expert manpower can be substituted by non-expert crowds, as exemplified by the numerous efforts to use Amazon Mechanical Turk (AMT) to build NLP datasets; e.g., Ganbold et al. (2018) or Potthast et al. (2018). The synergy exploited by this paradigm between NLP and CALL is particularly interesting. Indeed, NLP can be used to enhance CALL methods and, as such, it can grant an intrinsic added value for the target crowd that is not limited by any type of resource, unlike other crowdsourcing approaches relying on extrinsic added values (e.g. monetary incentives in AMT). In addition, from the NLP perspective, the crowd of learners that could potentially be reached is immense. Accordingly, an unprecedented amount of data could theoretically be crowdsourced by exploiting such a synergy.

Nicolas et al. (2021) showed that linguistic knowledge about an entry of an NLP dataset can be obtained at expert quality level, provided that the same judgement is asked to a sufficient number of learners. This is mostly true when simple Boolean questions are used. Even linguistic judgements of inferior reliability (e.g. 70%) contribute to approaching statistical certainty about the right answer to Boolean questions.[1] This approach favors quantity over quality to meet its goals and can be used to produce new entries or to validate the existing ones in an NLP dataset.

V-trel is a vocabulary trainer that implements this paradigm to teach and crowdsource knowledge on semantic relations between words (Lyding et al., 2019; Rodosthenous et al., 2019; Nicolas et al., 2021). V-trel includes two types of questions. Open questions ask for words sharing a spe-

cific relation with a given one (e.g., "give me a synonym of x"). Closed questions show a pair of words and ask the Boolean question of whether they share a specific relation (e.g., "are x and y synonyms?"). From the crowdsourcing perspective, open questions are mostly intended to crowdsource additional knowledge (i.e. to crowdsource new candidate entries), whereas closed questions are designed to crowdsource judgements on the knowledge suggested in the open questions or already encoded in ConceptNet (i.e. to validate existing entries or new candidate entries).

As empirically observed, closed questions should elicit positive and negative answers from learners (Rodosthenous et al., 2020). Otherwise, when learners understand that the trainer tends to continuously expect the same answer (e.g., "yes"), they tend to give the same default answer mechanically, without producing meaningful judgments.

We aim at generating closed questions expecting both types of answers. We need to identify pairs of synonyms to produce questions eliciting a positive answer and word pairs sharing a semantic relation other than synonymy (e.g. antonymy) to elicit negative answers. We refer to them as *non-synonyms* in the rest of the paper. To elicit positive answers, we use synonyms, such as "house" and "home". To elicit negative answers, we need non-synonyms such as "good" and "bad". It is worth noting that we do not consider as non-synonyms pairs of unrelated words such as "house" and "dog"' because they do not share any kind of semantic relation. Questions eliciting negative answers generated from them would be of poor quality and would not pose any challenge to learners.

Since v-trel favours teaching over crowdsourcing to maintain its pool of users, closed questions designed to crowdsource knowledge from learners should be served on a low frequency. This implies the need to decide which questions can be used for teaching and which for crowdsourcing purposes. Hence, the need of a confidence score to divide the questions into the two sets.

Our method aims at replacing the closed questions generated from ConceptNet, whose expected answers have quality issues since ConceptNet is, as most NLP resources, an imperfect dataset and for which we cannot tell apart the questions that can be used for teaching and for crowdsourcing purposes. Even though the aggregation of answers crowdsourced from learners would solve

---

[1]While the results obtained tend to confirm the viability of the approach, many aspects remain unexplored; e.g., the difficulty of a question in the aggregation process.

these problems, we have a cold start issue similar to the chicken and egg paradox: the issues cannot be solved without offering the tool but the tool cannot be offered without solving the issues first.

3 Related Work

With respect to the automatic generation of language learning exercises, only little automatic generation is performed directly from NLP resources so far. Most efforts focus on exercises known as a "cloze" (deletion) test, where learners have to fill word gaps in a text (Hill and Simha, 2016; Katinskaia et al., 2018; Lee et al., 2019). The literature from the last four editions of the top-two venues concerned with using NLP for CALL[2] confirms that current efforts, aside from ours, are mostly dedicated to the generation of cloze exercises (Santhi Ponnusamy and Meurers, 2021), the modelling of the learner knowledge (Araneta et al., 2020), or the detection and/or correction of mistakes in written text (Üksik et al., 2021). Some preliminary efforts exist on the automatic generation of exercises from Finnish and Hungarian NLP resources.[3] Despite the relatively narrow nature of the exercises we aim at generating[4], our work represents one of the few efforts targeting the automatic generation of language learning exercises from NLP resources.

Since our method generates pairs of synonyms from an existing knowledge base, it shares common ground with approaches to build or extend similar datasets. In that respect, the state of the art is mostly concerned with the creation and curation of WordNets for which various semi- and fully-automatic techniques have been developed, especially for languages other than English. Following Vossen (1996), these methods can be categorised as using either a *merge* or an *expansion* approach or both. The merge approach employs monolingual resources to create a standalone WordNet and was adopted for EuroWordNet (Vossen, 1998), the Polish WordNet (Derwojedowa et al., 2008), the Norwegian WordNet (Fjeld and Nygaard, 2009) and the Danish WordNet (Pedersen et al., 2009). The expansion approach uses a source WordNet and translates its synsets into the target language. It was used to build MultiWordNet (Pianta et al., 2002), the Finnish WordNet (Linden and Carlson, 2010), the French WordNet WOLF (Sagot and Fišer, 2008), and to enhance a Persian WordNet (Mousavi and Faili, 2017; Mousavi and Faili, 2021).

Our method employs an expansion approach: it projects knowledge from other languages onto Italian, but it differs in three aspects. First, it relies on a different type of dataset: ConceptNet.[5] Second, the output is not a final product, but a "raw" dataset to be polished by crowdsourcing. Third, it aims at identifying both synonyms and non-synonyms, whereas the aforementioned methods are mostly concerned with synonyms only.

4 Generating Candidate Pairs

Our hypothesis is that if two non-Italian words are marked as synonyms in ConceptNet and such words are translations of a pair of Italian words, then the Italian words are synonyms with a high likelihood. For instance, the pair {*house*, *home*} in English with respect to {*casa*, *abitazione*} in Italian. The greater the number of such pairs of non-Italian words are identified (e.g., {*maison*, *logement*} in French, {*casa*, *vivienda*} in Spanish), the more likely the Italian words are to be synonymous. Hereafter, we refer to the number of pairs of non-Italian words projected onto an Italian pair as *Nb-projected-syn-pairs*.

At the same time, we assumed that the incorrect candidate pairs of synonyms generated would mostly constitute a valid set of candidate pairs of non-synonyms. As such, most candidate pairs would be used to tackle our specific use case.

We used this logic for all languages available in ConceptNet. In order to seamlessly add the data already available on Italian synonyms, we considered the Italian part of ConceptNet as describing just another non-Italian language. Hence, we considered all Italian words as translations of themselves in this "extra" language.

We extracted $84,602$ candidate pairs of Italian synonyms and randomly sampled and evaluated a subset of $1,120$ pairs to build a gold standard.

The annotation procedure started by reflecting the information in well-known online Italian dictionaries: Treccani, De Mauro, Gabrielli, Sabatini-Coletti, Rizzoli, and Virgilio.[6] When a candidate pair was not found in these dictionaries, an annotator studied the definitions of the two words and searched for a third word referenced as a synonym of both words in the pair. We only kept instances where the annotator showed a high confidence. In total, 515 were labeled as correct pairs and 485 as incorrect. We discarded 120 pairs. From the 1,000 annotated instances, 403 directly reflect the information of reference dictionaries, whereas 597 reflect the stand of the annotator.

By extrapolating the ratio observed in the gold standard, we estimate that 51.2% of the candidate pairs ($\sim 43.4k$ pairs) are indeed synonyms. In comparison, 19,906 Italian word pairs are marked as synonyms in ConceptNet. We randomly sampled and annotated 200 of them with the procedure used to build the gold standard. Our estimation that 84% of them ($\sim 16.7k$ pairs) are valid. Our set of candidate pairs of Italian synonyms is thus larger, but has lower quality. Using these pairs directly to generate closed questions eliciting a positive answer would thus defeat our goal of improving the quality of the closed questions.

5 Computing Confidence Scores

We aim at discriminating between instances intended to generate questions eliciting positive and negative answers, while discriminating questions used for teaching or crowdsourcing purposes. We relied on a binary classifier to flag candidate pairs as correct and incorrect instances of synonyms. The predictions are used to decide on the kind of answers to elicit —candidate pairs predicted as correct are used to generate questions eliciting a positive answer and vice-versa. The associated confidence scores are used to discriminate between questions used for teaching and for crowdsourcing. We used the aforementioned gold standard to train the classifier.

The features are the following. (1) The aforementioned *Nb-projected-syn-pairs* for each pair.

[6]https://www.treccani.it; https://dizionario.internazionale.it; https://www.grandidizionari.it/Dizionario_Italiano/; https://dizionari.corriere.it/dizionario_italiano/; https://dizionari.corriere.it/dizionario_sinonimi_contrari/; https://sapere.virgilio.it.

model	precision	recall	F_1
Random forest	**72.2**	62.4	67.0
Logistic regression	61.1	79.8	**69.2**
Random Tree	71.9	60.8	65.9
Baseline	51.2	**100.0**	67.7

Table 1: Leave-one-out cross-validation performance of three classifiers when identifying correct synonym pairs against an all-correct baseline.

(2) To distinguish the languages from which the projection of knowledge happened, we computed per language the size of each subset of pairs of non-Italian words projected onto the candidate pair (which, together, sum up to *Nb-projected-syn-pairs*). (3) To express "relatedness", we computed the size of the set of non-Italian pairs of words both marked as sharing a semantic relation (i.e. not only synonymy) and as translations of the candidate pair. We refer hereafter to this number as *Nb-projected-all-pairs*. We also computed a ratio obtained by dividing *Nb-projected-syn-pairs* by *Nb-projected-all-pairs*. (4) To indicate if a candidate pair might be better suited to another semantic relation, we computed per semantic relation the size of the subsets of non-Italian pairs of words both marked as sharing a semantic relation other than synonymy and as translations of the candidate pair, as well as a ratio value by dividing these sizes by *Nb-projected-syn-pairs* and a difference by subtracting *Nb-projected-syn-pairs* to them. (5) A last set of features represents the most found relation, besides synonymy, in these non-Italian pairs of words (i.e. the top "competitor") by providing its type and duplicating the corresponding size of the subset of non-Italian pairs of words, ratio, and difference.

Since our gold standard is small, we ran a leave-one-out cross validation process to assess the quality of the predictions for a number of classifiers with default settings.[7] Table 1 shows the performance obtained by three of them plus a baseline that labels all pairs as correct. Even if the logistic regressor obtains the highest F_1, we adopt the model with the highest precision: the random forest. The reason is that we have observed empirically that precision is the most adequate indicator of how much the confidence scores would corre-

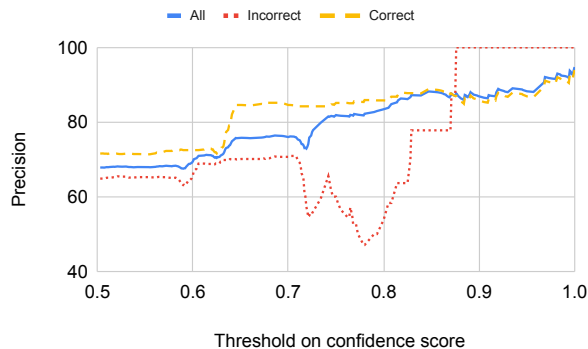[7]We used Weka 3.8.8; https://www.cs.waikato.ac.nz/ml/weka/.

Figure 1: Precision against confidence score threshold (lower-score predictions neglected).
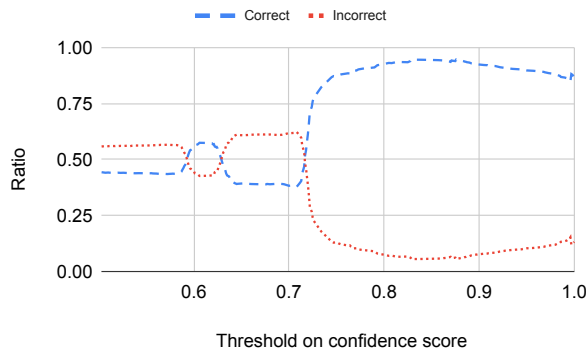

Figure 2: Ratio of label predicted according to thresholds on confidence scores (lower-score predictions neglected).

late with the quality of the predicted labels.[8]

6 Categorising Candidate Pairs

Once the binary classification was completed, we had to distinguish which pairs could be used to generate teaching and which to generate crowdsourcing questions. For that, we studied the correlation between confidence scores and quality of prediction.

Figure 1 shows the precision obtained when thresholding at different confidence score values. The "all" curve shows a clear correlation between the quality of the label predicted and the confidence scores, which was the main result we were aiming for. However, the performance differs noticeably with respect to the label predicted: the curve associated with pairs predicted as "correct" grows as expected, whereas the one for pairs predicted as "incorrect" does not. The reason can be observed through the ratio of labels predicted according to confidence scores.

As Figure 2 shows, label "incorrect" was rarely predicted with high confidence scores. This is because our method is inherently oriented towards identifying pairs of synonyms. Accordingly, the pairs outputted that are not synonyms are also not, as we hoped for, pairs of non-synonyms. They are mostly random noise induced by homonyms in other languages. For example, the candidate {fuoco, licenziare} was generated because the English words {fire, dismiss} are synonyms. Fire has several homonyms with different senses, one of which translates to fuoco in Italian. Our set of candidate pairs thus contains

only a few non-synonym pairs that the binary classifier struggles to spot. Therefore, our method cannot be used at present to generate closed questions eliciting negative answers.

This is not the case for pairs predicted as correct. For example, by using a minimum threshold of 0.996 on the confidence scores, we can select 3,829 pairs for which the predicted "correct" label is 94.44% reliable. This represents a set of pairs of reasonable size and better quality than the ones encoded in ConceptNet, which allow us to address part of the cold-start issue.

7 Conclusions and Ongoing Work

We presented a method to generate candidate pairs of Italian words that are synonyms or non-synonyms of one another from ConceptNet. These pairs will be used to generate questions used by a vocabulary trainer designed to combine the crowdsourcing of NLP datasets with language learning. While overtime all questions will be used for both teaching and crowdsourcing purposes, part of the pairs generated will at first be used to teach learners while the other part will at first be used to crowdsource knowledge in order to enhance ConceptNet. The obtained pairs, known to be correct synonyms in advance, can be served to the learners to improve their vocabulary skills. Another subset, whose correctness is still to be confirmed, can be served to the learners for validation and to decide whether the synonym connection between them should be added to ConceptNet or not.

Our results show that we can produce adequate data to generate part of the questions, while we are still unable to produce the data required to

---

[8]Future efforts will explore more direct and quantifiable means of formally informing this selection; cf. Section 7.

generate the complementary set of questions. In order to tackle the latter, we are devising a similar approach to identify candidate pairs of non-synonyms. We are adapting our overall procedure for the pairs of Italian words marked as translations of non-Italian words sharing any semantic relations (e.g. antonyms or hyponyms) instead of only considering the ones marked as translations of non-Italian words sharing a synonymy relation.

We are also interested in exploring possibilities to perform a more informed selection of the binary classification algorithm and will explore metrics to quantify the correlation between confidence scores and the quality of the predicted labels (e.g. Pearson, Kendall). In the future, we aim at running a crowdsourcing experiment with students of Italian as a second language with the produced data.

References

Marianne Grace Araneta, Gülşen Eryiğit, Alexander König, Ji-Ung Lee, Ana Luís, Verena Lyding, Lionel Nicolas, Christos Rodosthenous, and Federico Sangati. 2020. Substituto – a synchronous educational language game for simultaneous teaching and crowdsourcing. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Gothenburg, Sweden, November. LiU Electronic Press.

Magdalena Derwojedowa, Maciej Piasecki, Stanisªaw Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of polish wordnet. *Proceedings of GWC 2008*, pages 162–177.

Ruth Vatvedt Fjeld and Lars Nygaard. 2009. Nornet-a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources-between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7, pages 13–16.

Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using crowd agreement for wordnet localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30.

Anisia Katinskaia, Javad Nouri, Roman Yangarber, et al. 2018. Revita: a language-learning platform at the intersection of its and call. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ji-Ung Lee, Erik Schwan, and Christian M Meyer. 2019. Manipulating the difficulty of c-tests. *arXiv preprint arXiv:1906.06905*.

K Linden and L Carlson. 2010. Construction of a finnwordnet. *Nordic Journal of Lexicography*, 17:119–140.

Verena Lyding, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Lionel Nicolas, Alexander König, Jolita Horbacauskiene, and Anisia Katinskaia. 2019. v-trel: Vocabulary trainer for tracing word relations - an implicit crowdsourcing approach. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 674–683, Varna, Bulgaria, September. INCOMA Ltd.

Zahra Mousavi and Heshaam Faili. 2017. Persian wordnet construction using supervised learning. *arXiv preprint arXiv:1704.03223*.

Zahra Mousavi and Heshaam Faili. 2021. Developing the persian wordnet of verbs using supervised learning. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–18.

Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forăscu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, et al. 2020. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 268–278.

Lionel Nicolas, Lavinia Nicoleta Aparaschivei, Verena Lyding, Christos Rodosthenous, Federico Sangati, Alexander König, and Corina Forascu. 2021. An experiment on implicitly crowdsourcing expert knowledge about Romanian synonyms from language learners. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 1–14, Online, May. LiU Electronic Press.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a large corpus of

clickbait on twitter. In *Proceedings of the 27th international conference on computational linguistics*, pages 1498–1507.

Christos T. Rodosthenous, Verena Lyding, Alexander Koenig, Jolita Horbacauskiene, Anisia Katinskaia, Umair ul Hassan, Nicos Isaak, Federico Sangati, and Lionel Nicolas. 2019. Designing a prototype architecture for crowdsourcing language resources. In *LDK*.

Christos Rodosthenous, Verena Lyding, Federico Sangati, Alexander König, Umair ul Hassan, Lionel Nicolas, Jolita Horbacauskiene, Anisia Katinskaia, and Lavinia Aparaschivei. 2020. Using crowdsourced exercises for vocabulary training to expand conceptnet. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 307–316.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *OntoLex*.

Haemanth Santhi Ponnusamy and Detmar Meurers. 2021. Employing distributional semantics to organize task-focused vocabulary learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 26–36, Online, April. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Tiiu Üksik, Jelena Kallas, Kristina Koppel, Katrin Tsepelina, and Raili Pool. 2021. Estonian as a second language teacher's tools. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 130–134, Online, April. Association for Computational Linguistics.

Piek Vossen. 1996. Right or wrong: Combining lexical resources in the eurowordnet project. In *Proceedings of the 7th EURALEX International Congress*, pages 715–728, aug.

Piek Vossen. 1998. Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.