# Guessing the Age of Acquisition of Italian Lemmas through Linear Regression

**Irene Russo**
Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa
Via G. Moruzzi, 1, Pisa, Italy
irene.russo@ilc.cnr.it

## Abstract

The age of acquisition of a word is a psycholinguistic variable concerning the age at which a word is typically learned. It correlates with other psycholinguistic variables such as familiarity, concreteness, and imageability. Existing datasets for multiple languages also include linguistic variables such as the length and the frequency of lemmas in different corpora.

There are substantial sets of normative values for English, but for other languages, such as Italian, the coverage is scarce. In this paper, a set of regression experiments investigates whether it is possible to guess the age of acquisition of Italian lemmas that have not been previously rated by humans. An intrinsic evaluation is proposed, correlating estimated Italian lemmas' AoA with English lemmas' AoA. An extrinsic evaluation - using AoA values as features for the classification of literary excerpts labeled by age appropriateness - shows how essential is lexical coverage for this task.

## 1 Introduction

The age of acquisition (AoA, henceforth) of a word as the age at which a word was typically learned is a well-know psycholinguistic variable investigated for multiple languages (Moors et al., 2013; Ferrand et al., 2008; Alonso et al., 2012). It correlates with other variables such as concreteness of a word, frequency in a corpus, length of the word in letters and syllables. AoA estimates can be obtained by asking parents to record data about their children while they grow up. They are more frequently obtained by requesting the experiment's participants to indicate at which age they learned different words.

The collection of such ratings in an experimental setting is time-consuming. For this reason, existing datasets tend to have low coverage. With the advent of crowdsourcing platforms, there have been multiple efforts to enlarge the lists of AoA estimates asking for words' ratings. Participants self-assessed the age (in years) at which they thought they had learned the word, meaning that they would have understood that word even if they would not have been able to use, read, or write it. Crowdsourced AoA estimates show a good correlation with analogous ratings from traditional experimental settings, validating this methodology to improve AoA lists' coverage in less time (Kuperman et al., 2012).

For psycholinguists, the AoA of a word is a crucial variable in the selection process of the stimuli for lexical decision task experiments. Its correlation with other variables indicates that multiple factors should be taken into account when testing hypotheses involving lexical semantics, to ensure the right level of variability and complexity in the stimuli set.

These multiple factors correlated with AoA of a lemma make it possible to use a regression model to guess lemmas' AoA[1]. In this paper, a set of regression experiments investigates which features help to guess the age of acquisition of Italian lemmas that humans have not previously rated. An intrinsic evaluation is proposed, correlating estimated Italian lemmas' AoA with English lemmas' AoA, as proposed by (Montefinese et al., 2019) to test the generalizability of AoA ratings.

The extrinsic evaluation investigates if a list of AoA with better coverage could be beneficial for all those NLP tasks that classify or order texts according to their cognitive complexity, following the research hypotheses investigated for English (Xia et al., 2016; Vajjala and Meurers, 2014). More specifically, through a regression-as-classification approach, preliminary results on the role of AoA of lemmas for the classification of children's literature excerpts labeled with age appropriateness

[1] In lexicographic terms, this paper investigates the age of acquisition of a lemma since existing datasets record this feature for lemmatized word and not for their inflected forms.

*Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 43–48
Online Event, November 19, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

information are proposed.

## 2 Related Works

Several studies analyze how to extract psycholinguistic variables with corpus-based methodologies, with the aim of testing if this kind of knowledge is implicitly contained in the language. One practical aim is to create larger datasets of psycholinguistic norms avoiding the time-consuming phase of collections involving participants.

The seminal work of (Bestgen and Vincze, 2012) proposes latent semantic analysis to estimate lexical norms of words, obtaining satisfactory results for concreteness, imagery, and valence but less good for arousal and dominance. The age of acquisition is not included among the variables.

Mandera et al. (2015) build a semantic similarity space and apply machine learning techniques to extrapolate existing ratings of previously unrated words for five psycholinguistic properties (age of acquisition, concreteness, arousal, dominance and valence). The results are encouraging in terms of correlation with human norms. For example, the Pearson correlation is 0.737 for the age of acquisition of words, meaning that words related to similar topics are more likely to be acquired around the same age. However, according to the authors, evaluating the results in a lexical decision task is not satisfactory from a psycholinguistic perspective. The methodology used may introduce artifacts to the data and produce results that could lead to different conclusions that would be reached based on human ratings.

Hollis and Westbury (2012) use principal component analysis to understand the semantic dimensions along which the skip-gram model organizes meaning, finding how many dimensions are correlated with a particular semantic and lexical variable among the ones relevant for psycholinguistic research. Their findings confirm that the age of acquisition is in some way encoded in the semantic vector representation for word meanings, but no predictive methodology to assign a value to unrated words is implemented.

Vankrunkelsven et al. (2018) compare a distributional semantic model derived from word co-occurrences and a word association based model (with data collected from human subjects) in predicting psycholinguistic properties of words that affect lexical processing for Dutch. Overall, the correlations show a better performance of the methodology implemented (kNN) when word association information is used. Among the variables, estimates for the age of acquisition display the lowest correlation, while affective variables such as valence and concreteness are clearly encoded in the semantic models tested.

## 3 Datasets

A short description of all the datasets used in the experiments reported in Section 3 is provided in this section.

**Montefinese et al. (2019)**: this dataset contains ratings for 1,957 Italian content lemmas evaluated by 507 native Italian speakers recruited online. The stimuli were distributed over 20 lists containing 97–98 words each; each lemma was rated by 25 participants. The collected judgments are plausible because of strong internal reliability and a good correlation (Pearson r = 0.697) with translated English norms (Kuperman et al., 2012).

This dataset also contains information about the word length (i.e., number of characters), the word frequency from two different corpora (La Repubblica and ItWac), and the number of orthographic neighbors. Table1 reports the composition of the norms in terms of parts of speech.

| part of speech | #lemmas |
|---|---:|
| noun | 1494 |
| adjective | 311 |
| verb | 152 |

Table 1: Composition of (Montefinese et al. 2019) Italian AoA norms.

**Kuperman et al. (2012)**: this dataset - the biggest one available - is composed of 30,121 English words with AoA ratings obtained through crowdsourcing (e.g., through Amazon Mechanical Turk). The ratings are reliable as those obtained in laboratory conditions. In Table 2 its composition in terms of the parts of speech included in the Italian dataset (adverbs are excluded) is presented.

| part of speech | #lemmas |
|---|---:|
| noun | 18825 |
| adjective | 7259 |
| verb | 3622 |

Table 2: Composition of (Kuperman et. al 2012) English AoA norms.

**megahr_it**: this dataset contains concreteness and imageability estimates for 77 languages obtained through cross-lingual transfer via word embeddings (Ljubešić et al., 2018). Concreteness refers to the degree to which a concept denoted by a word refers to a perceptible entity. Imageability is a psycholinguistic variable that indicates how well a word gives rise to a mental image or sensory experience. The Italian dataset contains 100,000 words, 53% of them occurring among the most frequent 30,000 lemmas in La Repubblica lemmas' frequency list.

**La Repubblica lemmas' frequency list**: this list contains frequencies of lemmas in La Repubblica corpus (Baroni et al., 2004) and it's one the source for frequencies information included in (Montefinese et al., 2019).[2]

**Visual Genome lemmas' frequency list**: the Visual Genome dataset (Krishna et al., 2017) is the largest dataset of image descriptions for English. It is composed of dense annotations of objects, attributes, and relationships between objects for 108K images. As a pre-processing step, the descriptions have been annotated with TreeTagger (Schmid, 1994) and extracted the list of lemmas ordered by frequency. Frequencies in Visual Genome are included as a feature in linear regression experiments for estimating AoA of Italian lemmas to counterbalance La Repubblica frequency list where abstract meanings are more frequent.

## 4 Experiments

The training set is a subset of the Italian AoA norms dataset (Montefinese et al., 2019) resulting from the intersection of all datasets that contain features used for the linear regression experiments. As a consequence, it is smaller than the original dataset (see Table 3.)

| **part of speech** | **#lemmas** |
|---|---|
| noun | 1161 |
| adjective | 211 |
| verbs | 536 |
| **total** | 1908 |

Table 3: Composition of the training set.

The following features are included in the training set:

- L: lenght of each lemma;
- f_rep: the natural logarithm of the written frequency of lemmas in ”La Repubblica” corpus (Baroni et al., 2004);
- f_vg: the natural logarithm of the frequency of lemmas in the Visual Genome descriptions corpus, mapped onto Italian lemmas through Open Multilingual Wordnet's alignment (Krishna et al., 2017);
- concreteness: rating about the perceptibility of a concept denoted by a lemma, extracted from the mega_hr dataset (Ljubešić et al., 2018);
- imageability: rating about the strenght of sensory experience associated with a concept, extracted from the mega_hr dataset (Ljubešić et al., 2018).

A linear regression model is implemented and its performance is evaluated through 10-cross fold validation on the Italian norms training set. The results for different combination of features are reported in Table 4. The first column reports the mean standard error, a common evaluation measure for regression experiments. The second reports Pearson correlation between the estimated AoA and the real one, contained in (Montefinese et al., 2019)'s dataset. Concerning parts of speech, adjectives show the best correlation (0.61) while nouns are more problematic (0.55) and verbs are in between (r = 0.58)[3]. The all features combination is then applied to the evaluation of a list of 2,783 not previously rated lemmas obtained considering the most frequent 8,000 lemmas in La Repubblica corpus and providing for each of them the part of speech and the

| **features** | **MSE** | **Pearson r** |
|---|---|---|
| L + f_rep | -1.46 | 0.32 |
| L + f_vg | -1.45 | 0.34 |
| L + f_rep + f_vg | -1.37 | 0.45 |
| conc + imag | -1.47 | 0.30 |
| all features | -1.23 | 0.58 |

Table 4: MSE and Pearson correlation between real and estimated AoA of Italian lemmas.

[2] The list is available at wacky.sslmit.unibo.it

[3] All the correlations reported in this paper are significant at the 0.05 level.

| features | Pearson r |
|---|---|
| L + f_rep | 0.242 |
| L + f_vg | 0.578 |
| L + f_rep + f_vg | 0.592 |
| conc + imag | 0.190 |
| all features | 0.515 |

Table 5: Pearson correlation between real English and estimated AoA of aligned Italian lemmas.

English translation found in the Open Multilingual Wordnet (Bond and Paik, 2012).

The estimaed AoA ratings are evaluated through a comparison with the English ones provided by (Kuperman et al., 2012). This comparison represents an intrinsic evaluation of the models since Pearson correlation between Italian and English AoA lemmas has been used by (Montefinese et al., 2019) to validate the generalizability of the collected norms. The correlation reported by the authors was 0.697.

Table 5 reports the performance for new lemmas in terms of Pearson correlation with English lemmas. The best performance is achieved for verbs (0.646), then nouns (0.584) and adjectives (0.543). Surprisingly, the best result is not obtained with all features but with a combination of frequencies (from La Repubblica corpus and from the Visual Genome dataset) plus the length (i.e. number of character) of the lemmas.

## 5 Evaluation

The lists of 2,783 Italian lemmas with estimated AoA produced as a result of the linear regressions experiment presented in Section 3 can be evaluated on a dedicated dataset in a regression-as-classification task. A dataset of children's literature short texts is created, composed by epub excerpts made available by publishing houses[4]. From an e-commerce website[5] the appropriate age of potential readers is crawled.

The dataset is composed by 629 extracts (458,210 tokens in total, mean of each extract 728 tokens). Table 6 reports the composition of the children's literature excerpts corpus.

The set of features used for regression-as-classification experiments are based on the age of acquisition of lemmas:

- aoa_sum: sum of the age of acquisition values for rated lemmas as attested in texts;

| features | excerpts |
|---|---|
| from 8 years | 116 |
| from 9 years | 113 |
| from 10 years | 237 |
| from 11 years | 163 |

Table 6: Children's literature corpus, excerpts labeled by age appropriateness.

| features | accuracy |
|---|---|
| set 1 | 0.372 |
| set 2 | 0.354 |
| set 3 | 0.325 |
| set 4 | 0.348 |
| set 1 + set 3 | 0.335 |
| set 1 + set 4 | 0.330 |

Table 7: Pearson correlation between real and estimated age appropriateness of literary excerpts.

- aoa_mean: mean of the age of acquisition values in each text;
- aoa_std: standard deviation of the age of acquisition values in each text;
- aoa_max_value: maximum age of acquisition value occurring in a text;
- aoa_min_value: minimum age of acquisition value occurring in a text;
- max-min: difference between maximum and minimum age of acquisition values occurring in a text;
- frequency of occurrences from one to fourteen (set 2): number of occurrences of lemmas belonging to each age in the text;
- normalized frequency of occurrences from one to fourteen (set 3): number of occurrences of lemmas belonging to each age in the text, normalised by the total number of retrieved lemmas for each text;
- sum of all values from one to fourteen (set 4): sum of all values of lemmas belonging to each age in the text.

The best combination of features was found experimenting with (Montefinese et al., 2019)'s dataset, considering accuracy after rounding the linear regression outputs. In Table 7, the first six features constitute set 1.

[4] www.medialibrary.it

[5] www.ibs.it

| features | accuracy |
|---|---|
| L + f_rep | 0.378 |
| L + f_vg | 0.376 |
| L + f_rep + f_vg | 0.383 |
| conc + imag | 0.364 |
| all features | 0.379 |

Table 8: Pearson correlation between real and estimated age appropriateness of literary excerpts.

Since the best accuracy is obtained with features from set 1, the same set of features is applied for the evaluation of five AoA lists. Each list contains estimated AoA obtained with different set of features, as explained in Section 3. The aim is to test whether increasing the coverage of the Italian AoA dataset has positive effect on the classification of short texts by age appropriateness.

In line with what has been discovered about the correlation between English and Italian AoA values, the best set of features includes the frequencies from the two corpora and the lenght of the lemmas (see Table 8). Increasing (Montefinese et al., 2019)'s dataset with 2,783 lemmas with AoA estimated automatically slightly improves the accuracy in this specific classification task.

## 6 Conclusions and Future Works

In this paper, a set of regression experiments investigates if it is possible to guess the age of acquisition of Italian lemmas that humans have not previously rated by humans.

An intrinsic and extrinsic evaluation of the output is proposed. The results show that the overall quality of the estimated ratings enables their inclusion in NLP systems, even if they would not probably be satisfying for psycholinguistic experiments. More specifically, increasing the coverage of lexical resources containing AoA is beneficial for age appropriateness text classification.

As future work, the testing of semantic models for estimating the age of acquisition of Italian lemmas is relevant. Since the difficulty of a text could be assessed taking into account psycholinguistic variables that influence the cognitive complexity of the reading process, another interesting working hypothesis concerns the use of AoA features in other experiments involving the complexity of texts, such as readability assessment, L2 learners' written production, automatic assessment of text fluency for natural language generation outputs' evaluation.

## References

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2012. Subjective age-of-acquisition norms for 7,039 spanish words. *Behavior Research Methods*, (47):268–274.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the la repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper italian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yves Bestgen and Nadja Vincze. 2012. Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, (44):998–1006.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*.

Ludovic Ferrand, Patrick Bonin, Alain Méot, Maria Augustinova, Boris New, Christophe Pallier, and Marc Brysbaert. 2008. Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic french words and their relation with other psycholinguistic variables. *Behavior Research Methods*, (40):1049–1054.

Geoff Hollis and Chris Westbury. 2012. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin Review*, (23):1744–1756.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, (44):978–990.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia. Association for Computational Linguistics.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2015. How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly journal of experimental psychology*, (68(8)):1623–1642.

Maria Montefinese, David Vinson, Gabriella Vigliocco, and Ettore Ambrosini. 2019. Italian age of acquisition norms for a large set of words (itaoa). *Frontiers in Psychology*, 10:278.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Marteen De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic french words and their relation with other psycholinguistic variables. *Behavior Research Methods*, (45):169–177.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Sowmya Vajjala and Detmar Meurers. 2014. Exploring measures of “readability” for spoken language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 21–29, Gothenburg, Sweden. Association for Computational Linguistics.

Steven Vankrunkelsven, Verheyen, Gert Storms, and Simon De Deyne. 2018. Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition*, (1(1)).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.