Polarity Imbalance in Lexicon-based Sentiment Analysis

Marco Vassallo[1], **Giuliano Gabrieli**[1], **Valerio Basile**[2], **Cristina Bosco**[2]
1. CREA Research Centre for Agricultural Policies and Bio-economy, Italy
2. Dipartimento di Informatica, Università degli Studi di Torino, Italy
{marco.vassallo|giuliano.gabrieli}@crea.gov.it, {valerio.basile|cristina.bosco}@unito.it

Abstract

Polarity imbalance is an asymmetric situation that occurs while using parametric threshold values in lexicon-based Sentiment-Analysis (SA). The variation across the thresholds may have an opposite impact on the prediction of negative and positive polarity. We hypothesize that this may be due to asymmetries in the data or in the lexicon, or both. We carry out therefore experiments for evaluating the effect of lexicon and of the topics addressed in the data. Our experiments are based on a weighted version of the Italian linguistic resource MAL (Morphologically-inflected Affective Lexicon) by using as weighting corpus TWITA, a large-scale corpus of messages from Twitter in Italian. The novel Weighted-MAL (W-MAL), presented for the first time int this paper, achieved better polarity classification results especially for negative tweets, along with alleviating the aforementioned polarity imbalance.

Italiano. *Lo sbilanciamento della polarità è una situazione di asimmetria che si viene a creare quando si impiegano valori soglia parametrici nella Sentiment Analysis (SA) basata su dizionario. La variazione dei valori soglia può avere un impatto opposto rispetto alla predizione di polarità negativa e positiva. Si ipotizza che questo effetto sia dovuto ad asimmetrie nei dati o nel dizionario, o in entrambi. Abbiamo condotto esperimenti per misurare l'effetto del lessico e degli argomenti trattati nel nostro dataset. I nostri esperimenti sono basati su una versione ponderata della risorsa per l'italiano MAL (Morphologically-inflected Affective Lexicon), usando come corpus per la ponderazione TWITA, un corpus di larga scala di messaggi da Twitter in italiano. La nuova risorsa Weighted-MAL (W-MAL), presentata per la prima volta in questo articolo, ottiene migliori risultati nella classificazione della polarità specialmente, per i messaggi negativi, oltre ad alleviare il problema sopracitato di sbilanciamento della polarità.*

1 Introduction and Motivation

Sentiment Analysis (SA) is the task of Natural Language Processing that aims at extracting opinions from natural language expressions, e.g., reviews or social media posts. The basic approaches to SA typically fall into one of two categories: dictionary-based and supervised machine learning. Methods based on a dictionary make use of *affective* lexicons, language resources where each word or lemma is associated to a score indicating its affective valence (e.g., *polarity*). In SA they are faster than supervised statistical approaches and require minimal adaptation, unless the resource is domain-specific, also when applied to multiple environments with minimal adaptation overhead. However, they only achieve good performance for identifying coarse opinion tendencies in large datasets, since they cannot take into account the impact of the context on the polarity value associated to a word.
Supervised statistical methods, on the other hand, tend to provide better quality predictions across benchmarks, due to their better ability to generalize over individual words and expressions, and learning higher level features. These models also show a better ability to adapt to specific domains,

provided the availability of data suitable for training.

In order to access the lexical entries in an affective dictionary, lemmatization must be performed on each single word. Unfortunately, lemmatization is an error-prone process, with potentially negative impact on the performance of downstream tasks such as SA. Vassallo et al. (2019) introduced a novel computational linguistic resource, namely the Morphologically-inflected Affective Lexicon (henceforth MAL) in order to address this issue by avoiding the lemmatization step in favor of a morphologically rich affective resource.

In the experiments we carried out on a specific text genre, namely social media, we have observed that using a threshold to assign polarity classes is beneficial, and using the MAL instead of a lemmatization step improves the SA performance overall, in particular due to a better prediction of the negative polarity. However, the variation in threshold has opposite impact on the prediction of negative and positive tweets.

In this paper, we investigate the motivation beyond this polarity imbalance. In particular, we speculate that this may be due to asymmetries in the data (e.g., different internal topics), in the lexicon (e.g., different amounts of negative and positive terms), or both, and we provide experiments to better understand this result and validate these hypotheses. We can therefore summarize as follows our research questions:

- Is the polarity imbalance due to the topic addressed?
- Is the polarity imbalance due to the lexicon (i.e., the resources we used, Sentix and MAL)?
- Is the polarity imbalance due to both?

A further contribution of the paper consists in providing a statistical method for finding the threshold for using the lexicon in SA tasks.

The paper is organized as follows. In the next section, affective lexicons and the resource MAL are discussed. In section 3, we describe the issues related to polarity imbalance in lexicon-based approaches for SA. The fourth section is instead devoted to discuss the impact on SA of lexicon and to introduce W-MAL. Section 5 discusses how the topics addressed in the text may impact ob SA. The final section provides conclusive remarks and some hints about future work.

2 Affective Lexicons

SA is typically cast as a text classification task, very often approached by supervised statistical models among the NLP research community (Barbieri et al., 2016). However, there are several scenarios where dictionary-based methods are preferred, including large-scale industry-ready systems, and domain-specific applications. While generally less accurate than supervised classification, dictionary-based methods tend to be robust to the classification of sentiment across different domains, faster and with a higher level of scalability.

For the Italian language, several sentiment dictionaries, or, using a more general term, *affective lexicons*, were published with different levels of granularity of the annotation and availability to the public, as summarized on the website of the Italian Association of Computational Linguistics[1].

Sentix (Basile and Nissim, 2013) is one of the first affective lexicons created for Italian language, with a first release described in (Basile and Nissim, 2013), and a second release called Sentix 2.0[2]. It provides an automatic alignment between SentiWordNet, an automatically-built polarity lexicon for English by Baccianella et al. (2010), and the Italian portion of MultiWordNet (Pianta et al., 2002). While the first version of Sentix associated two independent positive and negative polarity scores to each word, in Sentix 2.0[3] all the senses of each lemma have been collapsed into one entry by means of a weighted average, where the weights are proportional to sense frequencies computed on the sense-annotated corpus SemCor (Langone et al., 2004). Moreover, the positive and negative polarity scores have been combined to form a single polarity score ranging from -1 (totally negative) to 1 (totally positive). Sentix 2.0 includes 41,800 different lemmas.

In order to use a lemma-based affective lexicon such as Sentix, lemmatization is a necessary step to undertake. In our previous work, we found that such intermediate step causes a considerable amount of noise, in the form of lemmatization er-

[1] http://www.ai-lc.it/en/affective-lexica-and-other-resources-for-italian/

[2] https://github.com/valeriobasile/sentixR

[3] https://github.com/valeriobasile/sentixR

Table 1: A tweet with the output of the three lemmatization models where the lemmas are alphabetically ordered and the errors marked in bold.

Original	`@ANBI_Nazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgriculturaIT`
ISDT	`acqua` **`adigire`** `allarme crisi` **`d`** `dopo idrico po` - Sentix score: 0.080
POSTWITA	`acqua` **`adigere`** `allarme crisi di dopo idrico po` - Sentix score: 0.080
PARTUT	**`acquare adigere`** `allarme crisi` **`d`** `dopo idrico po` - Sentix score: -0.078

rors such as the ones shown in Table 1 (Vassallo et al., 2019). We terefore built a new resource on top of Sentix, described in the next section.

2.1 MAL

We proposed the Morphologically-inflected Affective Lexicon in Vassallo et al. (2019, MAL). It is an extension of Sentix where the entries associated to polarity scores rather than lemmas are the inflected forms related to each lemma, and the polarity scores to be associated to each form are drawn from the original lemmas in Sentix. The approach consists in linking the lexical items found in tweets with the entries of Sentix 2.0, without the application of an explicit lemmatization step. The lexicon is indeed expanded by considering all the acceptable forms of its lemmas extracted from the Morph-It collection of Italian forms (Zanchetta and Baroni, 2005). Each form takes the same polarity score of the original lemma, but when different lemmas can assume the same form, the arithmetic mean of their polarity scores is assigned. The MAL comprises 148,867 forms and all the items linked to the lemmas of Sentix 2.0 .

Using the MAL we performed a series of experiments on the impact of lemmatization on dictionary-based SA, which showed how the reduction in lemmatization errors leads to a better polarity classification performance.

3 Polarity Imbalance in Lexicon-based Sentiment Analysis

When using an affective lexicon to predict the polarity of natural language sentences, a threshold must be fixed to translate the numerical scores into discrete classes, e.g., positive, neutral, and negative. In Vassallo et al. (2019), we showed how the variation of such threshold has different, opposite impacts on the accuracy of the classification, using as a benchmark the corpus annotated with sentiment polarity made available by the SENTIment POLarity Classification (SENTIPOLC) shared task at EVALITA 2016. More precisely, the red dotted lines with label ALL in Figure 1 show that the F1 score of the classification of positive polarity instances increases with stricter thresholds, while the F1 score of negative polarity instances decreases.

We postulate two non-mutually exclusive hypotheses on the origin of the polarity imbalance, namely the effect of lexicon and topic. The affective scores in the lexicon may be biased towards one end of the polarity spectrum due to a number of causes, resulting in skewed classification results. On the other hand, some topics tend to attract opinions more polarized towards one end of the spectrum than the other (e.g., "war" is an inherently negative topic), therefore the classification might be influenced by this intrinsic polarization.

4 The Effect of Lexicon on SA

In order to shed some light on the polarity imbalance due to lexicon we applied a weighted approach to MAL by developing the Weighted Morphologically-inflected Affective Lexicon (W-MAL). It originates from the intuition that less frequent terms should have a higher impact on the computation of the polarity of the sentence where they occur. This principle stems from the observation that more sought-after terms are often used to convey stronger opinions and feelings.

We therefore computed the relative frequency of every item in MAL by using TWITA, a large-scale corpus of messages from Twitter in the Italian language (Basile et al., 2018). TWITA is indeed large (covering over 500 million tweets from 2012 to 2018, and the collection is currently ongoing) and domain-agnostic enough to provide a sufficiently representative sample of the distribution of the Italian language words, although specific to one social media platform.

Despite its size, not all the terms from the MAL occur in TWITA: 57.9% of the 148,867 terms occurring in MAL were found in TWITA, due to the sparseness of particular inflected forms, and to the presence of multi-word expressions in the lexicon (18,661, about 12%) that were not considered for

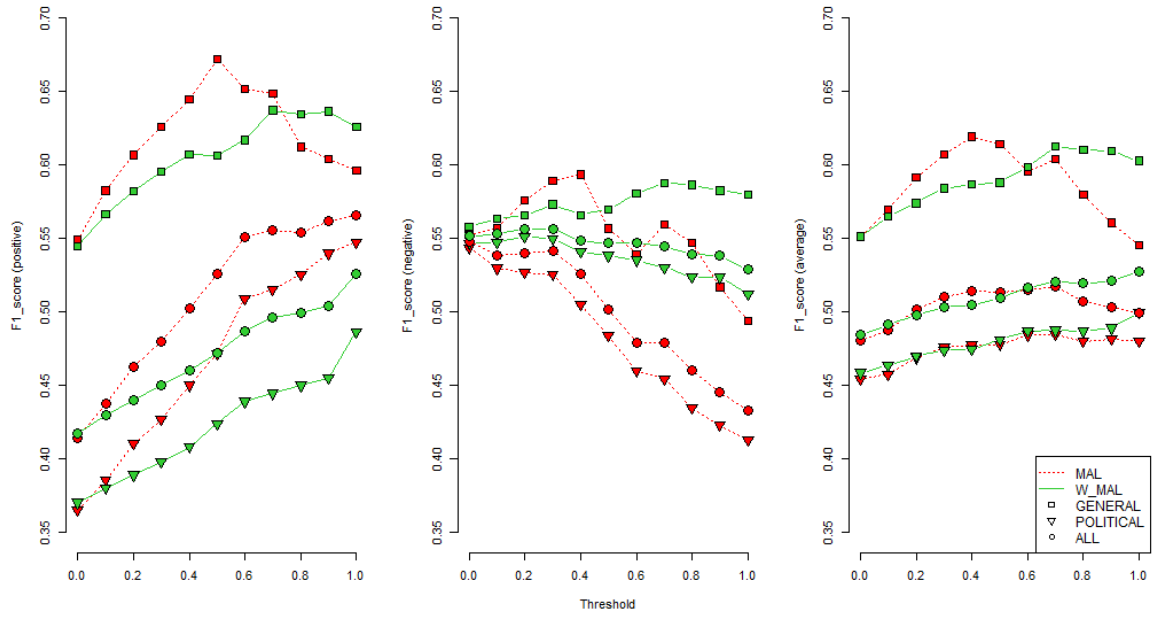

Figure 1: Results of the polarity classification on SENTIPOLC. The threshold value on the X-axis is applied to transform the sum of the scores from the lexicon into a positive or negative label.

matching the resources. For comparison, 73,36% of Sentix lemmas were found in TWITA.

Accordingly, the scores of MAL were recalculated by weighting them with the associated words frequency in TWITA, using the Zipf scale measure (van Heuven et al., 2014). We decided to use this measure because of its easy understanding and the short computation timing. Actually, the Zipf scale measure is a logarithmic scale based on the well-known Zipf law of word frequency distribution (Zipf, 1949). The computation of Zipf values of terms frequencies from TWITA is straightforward and essentially equals to the logarithm of the absolute frequency scaled down by a multiplicative factor:

$$Zipf(i) = log_{10}\left(\frac{f(i)}{\frac{\sum_{i=1}^{N} f(i)}{10^6} + \frac{N}{10^6}}\right) + 3$$

where N is the number of tokens in TWITA (6,644,867), $f(i)$ is the absolute frequency of the $i-$th token in TWITA, and the sum of the token frequencies $\sum_{i=1}^{N} f(i) = 6,906,070,053$, therefore:

$$Zipf(i) = log_{10}\left(\frac{f(i)}{6,906.07 + 6.644}\right) + 3$$

The original Zipf scale is a continuous scale and it ranges from 1 (very low frequency) to 6 (very high frequency) or even 7 (e.g., for very frequent words like auxiliar verbs). By computing the Zipf score of the MAL terms on TWITA, we found some terms with very low frequencies, resulting in negative values because of the logarithmic function. These were re-coded with the minimun Zipf value. The resulting weights in the W-MAL range from a minimum of -5.16 to a maximum of 5.95 (the original MAL ranged from -1 to 1). Eventually, we decided to keep the terms that were not found in TWITA in the W-MAL with their MAL original score.

We initially applied the Zipf scale to MAL polarity scores by simply multiplying the two found scores and thus giving more weight to high frequent terms. However, using the affective lexicon with such weighting scheme resulted in a decrease in its polarity classification performance. We therefore simply reversed the Zipf scale by weighting the original scores inversely with respect to their words frequency. By doing so, we tested for our speculation of giving more weight to low frequent terms. We replicated the polarity detection experiment on SENTIPOLC. The results, shown in the green solid lines in Figure 1 labeled ALL, indicate a better performance over-

all, and a reduced imbalance between the positive (F1-scores standard deviation across the thresholds of 0.035 with W-MAL vs 0.054 with MAL) and (especially) the negative polarity class (F1-scores standard deviation across the thresholds of 0.008 with W-MAL vs 0.042 with MAL).

To further clarify the effect found on the polarity scores, we show two example tweets in Figure 2[4]. In the figure, the MAL and W-MAL scores are included for the highlighted words, along with the total polarity scores computed with both dictionaries, showing how the final judgment can change from neutral to polarized (bottom example) or switch polarity entirely (top example). In particular in the top example the scores are associated with "confondesse" (to confuse in subjunctive mood) and to "diritto" (right), while in the bottom example the scores are associated with "Istituto" (school) and to the periphrastic verbal form "viene taciuto" (is silenced). This result confirms our speculation that negative polarity is expressed with more specific words than positive polarity. Psychology studies also show that more complex forms of language were used for expressing criticisms rather that positive evaluations (Stewart, 2015).
We also notice how the F1-score on the negative polarity is generally higher than the one on the positive polarity class. This means that the negative polarity of tweets is better predicted than the positive polarity by means of the weighted process with the inverse coding. This outcome seems to be substantially supported also by the W-MAL directly proportional performance that worked worse than the inverse version in terms of prediction. This trend was also observed across most of the results of the SENTIPOLC shared task, mostly based on supervised models with lexical features, further indicating that the vocabulary of negative sentiments is richer than that of positive sentiment.

[4]The translation of the examples is as follows. For the top example: *They would be #thegoodschool if meritocracy were not confused with "doormatcracy": the one whereby even a right becomes a concession.* For the bottom example: *@steGiannini #thegoodschool In the rankings of the School there are also TFA qualified teachers with 48 months of service. Why is it silenced?* where steGiannini refers to the Italian minister for school

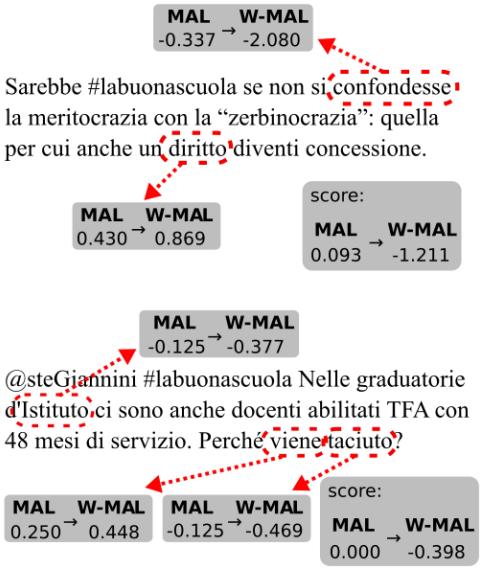

Figure 2: A comparison between the scores calculated for polarized words of a tweet according to MAL and W-MAL in two tweets from the test set.

5 The Effect of Topic on Sentiment Analysis

In order to investigate the interaction between the imbalance of dictionary-based polarity classification and a possible asymmetry in the data (i.e. different internal topics), we performed such classification with MAL and W-MAL with the reversed Zipf scale on a benchmark with explicitly stated topics. As a matter of fact, the test set of SENTIPOLC is composed of 1,982 Italian tweets, organized in 496 *general* i.e. domain-independent tweets, and 1,486 *political* tweets, obtained by filtering data with specific keywords related to political Italian figures. The results of our experiment are also included in Figure 1 with the GENERAL and POLITICAL labels.

The first observation we draw from this experiment is that the polarity imbalance is a phenomenon restricted to the topic-specific section of the dataset. This confirms the hypothesis that dictionary-based polarity classification is affected by the imbalance issue with the extent to which its topic is specific. In particular, we hypothesize that some topics (such as politics) tend to attract opinions more polarized towards one end of the spectrum (the negative one in this case), therefore inducing the observed imbalance.

The second observation is that weighting the polarity scores in the dictionary based on word frequency (W-MAL) provides better overall results.

In particular, the F1 scores are better in the topic-specific case, specifically due to a better prediction of the negative polarity. This result reinforces the idea that a polarized topic induces polarity imbalance, and therefore a method to alleviate such imbalance (i.e., a weighting scheme) leads to better performance. In our view, a reason for this effect is that topic-specific messages make use of less frequent words on average.

6 Conclusion and Future Work

The weighting scheme proposed in this work is a promising solution to the polarity imbalance in dictionary-based SA. The experiments show that weighting the polarity scores with word frequencies yielded a more precise prediction of the polarized tweets, with lessened bias in the thresholds for neutral scores. The novel resource here presented, W-MAL, is an attempt to better characterize the most sought-after words, which have an impact on the interaction between sentiment and topic. We believe it also represents a promising attempt to control for context-dependency while using lexicon-based methods for SA.

In particular, with this resource we try to give voice to the linguistic intuition that the exploitation of a specific form within a message might meaningfully impact on the sentiment expressed in the message. For instance, referring to the top example in figure 2, by exploiting the subjunctive mood "confondesse" of the verb "confondere" (to confuse), the author joins together with the meaning of the verb also a sense of doubtfulness and of unreality. This is also improved by the fact that this form introduces a clause which is coordinated with the clause headed by a verb in conditional mood, i.e. "sarebbe" (form of to be). This form of the verb "confondere" seems especially adequate for contexts where a negative polarity is expressed and less appropriate for other cases. The use of this specific mood for the verb has therefore a meaningful impact on the sentiment expressed. The MAL properly encodes this information, which may be lost when a lemmatization step is applied on text and all forms are subsequently considered as bearing the same meaning without further nuances. But the W-MAL does also better: it encodes the probabilistic information about how suitable a form is for expressing a particular sentiment with respect to other available forms in a given context.

For all the aforementioned reasons, this work has drawn our attention to the necessity of weighting the dictionary-based affective lexicons to SA with corpora-based word frequencies. The resource is freely available at `https://github.com/valeriobasile/sentixR/blob/master/sentix/inst/extdata/W-MAL.tsv`

In future work, we plan on working on more refined weighting strategies, e.g., leveraging the frequency information of word forms in addition to lemmas, and taking the topic distribution into consideration. Reducing the computation load is a challenging goal as well (see Prakash et al. (2015)). On the other hand, modern transformer-based models have reached state-of-the-art results on the task of polarity detection (Polignano et al., 2019), although they are far more expensive and time- consuming to run. We plan therefore to compare the predictions of these systems, and study ways to integrate their respective strengths (i.e., speed and transparency of the dictionary-based approach vs. the superior prediction capability of the deep neural models) in order to boost the overall performance.

The present work was originally conceived in the framework of the AGRItrend project led by the CREA Research Centre for Agricultural Policies and Bio-economy, aiming at collecting and analyzing social media data for opinions in the domain of public policies and agriculture. As such, we plan on studying the impact of the techniques presented in this paper on that particular domain, and observe if the same, or different, patterns emerge. On a similar line, so far we conducted experiments on data from Twitter, which facilitates access to large quantity of data but restricts the range of text style and genre found in them.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment

POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR-WS.org.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. CEUR-WS.org.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 63–69. Association for Computational Linguistics (ACL).

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR-WS.org.

Saurabh Prakash, T. Chakravarthy, and E. Kaveri. 2015. Statistically weighted reviews to enhance sentiment classification. *Karbala International Journal of Modern Science*, 1:26–31.

Martyn Stewart. 2015. The language of praise and criticism in a student evaluation survey. *Studies In Educational Evaluation*, 45:1–9.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67:6:1176–1190.

Marco Vassallo, Giuliano Gabrieli, Valerio Basile, and Cristina Bosco. 2019. The tenuousness of lemmatization in lexicon-based sentiment analysis. In *Proceedings of the Sixth Italian Conference on Computational Linguistics - CLiC-it 2019*. Academia University Press.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).

George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley.