Quantifying the Semantic Core of Gender Systems

Adina Williams[ə] **Ryan Cotterell**[ʃ,ɦ] **Lawrence Wolf-Sonkin**[ʃ]
Damián Blasi[ʔ] **Hanna Wallach**[3]
[ə] Facebook AI Research, New York, USA
[ʃ]Department of Computer Science, Johns Hopkins University, Baltimore, USA
[ɦ]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[ʔ]Comparative Linguistics Department, University of Zürich, Switzerland
[3]Microsoft Research, New York City, USA
adinawilliams@fb.com, rdc42@cam.ac.uk, lawrencews@jhu.edu
damian.blasi@uzh.ch, hanna@dirichlet.net

Abstract

Many of the world's languages employ grammatical gender on the lexeme. For example, in Spanish, the word for house (*casa*) is feminine, whereas the word for paper (*papel*) is masculine. To a speaker of a genderless language, this assignment seems to exist with neither rhyme nor reason. But is the assignment of inanimate nouns to grammatical genders truly arbitrary? We present the first large-scale investigation of the arbitrariness of noun–gender assignments. To that end, we use canonical correlation analysis to correlate the grammatical gender of inanimate nouns with an externally grounded definition of their lexical semantics. We find that 18 languages exhibit a significant correlation between grammatical gender and lexical semantics.

1 Introduction

In his semi-autobiographic work about his time traveling through Germany, *A Tramp Abroad*, Twain (1880) recounted his difficulty when learning the German gender system: "Every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl." Although this humorous take on German grammatical gender is clearly a caricature, the quote highlights the fact that the relationship between the grammatical gender of nouns and their lexical semantics is often quite opaque.

As arbitrary as certain noun–gender assignments may appear overall, a relatively clear relationship often exists between grammatical gender and lexical semantics for some of the lexicon. The portion of the lexicon where this relationship is clear usually consists of animate nouns; nouns referring to people morphologically reflect the sociocultural notion of "natural genders." This portion of the lexicon—the "semantic core"—seems to be present in *all* gendered languages (Aksenov, 1984; Corbett, 1991). But how many *inanimate* nouns can also be included in the semantic core? Answering this question requires investigating whether there is a correlation between grammatical gender and lexical semantics for inanimate nouns.

Our primary technical contribution is demonstrating that grammatical gender and lexical semantics can be correlated using canonical correlation analysis (CCA)—a standard method for computing the correlation between two multivariate random variables. We consider 18 gendered languages, following 3 steps for each: First, we encode each inanimate noun as a one-hot vector representing the noun's grammatical gender in that language; we then create 5 operationalizations of each noun's lexical semantics using word embeddings in 5 genderless "donor" languages (English, Japanese, Korean, Mandarin Chinese, and Turkish); finally, for each genderless language, we use CCA to compute the desired correlation between grammatical gender and lexical semantics. This process yields a single value for each of the 90 gendered–genderless language pairs, revealing a significant correlation between grammatical gender and lexical semantics for 55 of these language pairs.

Secondarily, we investigate semantic similarities between the 18 languages' gender systems—i.e., their assignments of nouns to grammatical genders. We analyze the projections of lexical semantics (operationalized as word embeddings in English) obtained via CCA, finding that phylogenetically

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5734–5739, Hong Kong, China, November 3–7, 2019.

similar languages have more similar projections.

2 Background and Assumptions

2.1 Grammatical Gender

Languages range from employing no grammatical gender on inanimate nouns, like English, Japanese, Korean, Mandarin Chinese, and Turkish, to drawing grammatical distinctions between tens of gender-like classes (Corbett, 1991). Although there are many theories about the assignment of inanimate nouns to grammatical genders, to the best of our knowledge, the linguistics literature lacks any large-scale, quantitative investigation of arbitrariness of noun–gender assignments. However, with the advent of modern NLP methods—particularly with advancements in distributional approaches to semantics (Harris, 1954; Firth, 1957)—and with the copious amounts of text available on the internet, it is now possible to conduct such an investigation. We focus on languages that have either two (masculine–feminine) or three (masculine–feminine–neuter) genders, to which nouns are exhaustively assigned, and investigate whether a correlation exists between grammatical gender and lexical semantics for inanimate nouns—i.e., whether noun–gender assignments are arbitrary or not.

In many languages, a noun's grammatical gender can be predicted from its spelling and pronunciation (Cucerzan and Yarowsky, 2003; Nastase and Popescu, 2009). For example, almost all Spanish nouns ending in *-a* are feminine, whereas Spanish nouns ending in *-o* are usually masculine. These assignments are non-arbitrary; indeed, Corbett (1991, Ch. 4) provides a thorough typological description of how phonology pervades gender systems. We emphasize that these assignments are not the subject of our investigation. Rather, we are concerned with the relationship between grammatical gender and lexical semantics—i.e., when asking why the Spanish word *casa* is feminine, we do not consider that it ends in *-a*.

Finally, our investigation is related to that of Kann and Wolf-Sonkin, which assumes that noun–gender assignments are non-arbitrary and examines the predictability of grammatical gender from lemmatized word embeddings; in contrast, we investigate the arbitrariness of noun–gender assignments.

2.2 Lexical Semantics via Word Embeddings

The NLP community has widely adopted word embeddings as way of representing lexical semantics. The underlying motivation behind this adoption is the observation that words with similar meanings will be embedded as vectors that are closer together. As we explain in §3, our investigation requires a definition of lexical semantics that is independent of grammatical gender. However, in many gendered languages, word embeddings effectively encode grammatical gender because this information is trivially recoverable from distributional semantics. For example, in Spanish, singular masculine nouns tend to occur after the article *el*, whereas singular feminine nouns tend to occur after the article *la*. For this reason, we use an externally grounded definition of lexical semantics: we create 5 operationalizations of each noun's lexical semantics using word embeddings in 5 genderless "donor" languages (English, Japanese, Korean, Mandarin Chinese, and Turkish). We use 5 languages that are phylogenetically distinct and spoken in distinct regions to minimize any spurious correlations.[1]

Our investigation is based on the linguistic assumption that word embeddings in a genderless "donor" language are a good proxy for genderless lexical semantics. In practice, however, this assumption is generally false: word embeddings are largely a reflection of the text with which they were trained. For example, the embedding of the word *snow* will differ depending on whether the training text was written by people near the equator or people near the North Pole, even if both groups speak the same language. Such differences will be more pronounced for rare words, which are arguably more language- and culture-specific than many common words. For this reason, we limit the scope of our investigation to only those inanimate nouns that are likely to be used consistently across different languages. To implement this limitation, we use a Swadesh list (Buck, 1949; Swadesh, 1950, 1952, 1955, 1971/2006)—a list of words constructed to contain only very frequent words that are as close to culturally neutral as possible. By limiting the scope of our investigation to only those inanimate nouns that appear in a Swadesh list, we can be reasonably confident that their word embeddings in English, Japanese, Korean,

[1] It is natural to ask whether polysemy and homonymy might result in spurious similarities. For example, in English, the words *fish*$_N$ and *fish*$_V$ are homonymous, but in Mandarin Chinese, the words *yú* (*fish*$_N$) and *diào* (*fish*$_V$) are not. If patterns of homonymy in the genderless languages are very different, but patterns of correlation between grammatical gender and lexical semantics are very similar, then we can be reasonably sure that the correlations are not due to homonymy.

	bg	ca	el	es	fr	he	hi	hr	it	lt	lv	pl	pt	ro	ru	sk	sl	uk
en	**2596**	**2720**	**2872**	**4947**	**6257**	**1489**	828	443	**4800**	**923**	**881**	**1646**	**3918**	397	**5779**	1056	293	975
ja	**2586**	**2596**	**2886**	**3383**	**4654**	**2223**	**1421**	486	**3849**	1241	**1215**	**1884**	**3615**	**497**	**4532**	375	419	1380
ko	**1856**	**1843**	1982	1840	**2812**	**1774**	1269	371	**2513**	1089	**997**	1357	**2442**	**364**	**2680**	247	317	1169
tr	**1578**	**1623**	**1735**	**1766**	**2580**	**1275**	817	274	**2287**	903	826	1163	**2223**	303	**2470**	218	281	909
zh	**2275**	**2190**	2454	**2693**	**3722**	1810	**1266**	373	**3170**	**1111**	**1110**	**1643**	**3084**	480	**3652**	286	**341**	1196

Table 1: The number of inanimate nouns for each gendered–genderless language pair. Bold indicates that our investigation reveals a significant correlation between grammatical gender and lexical semantics for that pair.

Mandarin Chinese, and Turkish are a good proxy for their genderless lexical semantics, as desired.

3 Methodology

3.1 Data

We use Open Multilingual WordNet (Stamou et al., 2004; Ordan and Wintner, 2007; Raffaelli et al., 2008; Sagot and Fišer, 2008; Tufiş et al., 2008; Piasecki et al., 2009; Simov and Osenova, 2010; Toral et al., 2010; Bond and Paik, 2012; Fišer et al., 2012; González-Agirre et al., 2012; de Paiva and Rademaker, 2012; Rudnicka et al., 2012; Bond and Foster, 2013; Garabík and Pileckytė, 2013; Oliver et al., 2015)[2] as our Swadesh list. This yields 18 gendered languages (Bulgarian, Catalan, Greek, Spanish, French, Hebrew, Hindi, Croatian, Italian, Lithuanian, Latvian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, and Ukranian), of which 12 are from the Indo-European family and 4 are not.[3] For the 5 genderless languages (English, Japanese, Korean, Mandarin Chinese, and Turkish), we use pre-trained, 50-dimensional word embeddings from FASTTEXT (Bojanowski et al., 2017; Grave et al., 2018).[4] For each gendered–genderless language pair, we limit the scope of our investigation to only those inanimate nouns that occur in both our Swadesh list and in FASTTEXT; we provide the resulting counts in Table 1. Finally, we randomly partition the set of nouns for each language pair into a 75%–25% training–testing split.

3.2 Notation

We first establish the requisite notation. Let $\mathcal{V}_{\ell,m} = \{1, \ldots, V_{\ell,m}\}$ denote a set of integers representing the inanimate nouns for gendered language ℓ and genderless language m. Let $\mathcal{G}_\ell$ denote the (arbitrarily ordered) genders in language ℓ; for example, let $\mathcal{G}_{spanish} = (\text{MSC}, \text{FEM})$. Given an inanimate noun $n \in \mathcal{V}_{\ell,m}$, let $\boldsymbol{g}_\ell(n)$ denote a one-hot vector representing n's grammatical gender in language ℓ, so that the i^{th} entry corresponds to the i^{th} gender in $\mathcal{G}_\ell$. Similarly, let $\mathbf{e}_m(n) \in \mathbb{R}^{50}$ denote the 50-dimensional word embedding representing the lexical semantics of n in language m. Let $G_\ell \in \mathbb{R}^{|\mathcal{G}_\ell| \times V_{\ell,m}}$ collectively denote the inanimate nouns' grammatical genders in language ℓ, so that the n^{th} column is $\boldsymbol{g}_\ell(n)$, and let $E_m \in \mathbb{R}^{50 \times V_{\ell,m}}$ collectively denote the inanimate nouns' lexical semantics, so that the n^{th} column is $\mathbf{e}_m(n)$. Finally, let G_ℓ^{train} and E_m^{train} respectively denote the columns of G_ℓ and E_m that correspond to the inanimate nouns in the training set and let G_ℓ^{test} and E^{test} respectively denote the columns of G_ℓ and E_m that correspond to the inanimate nouns in the testing set.

3.3 Canonical Correlation Analysis

CCA is a standard method for computing the correlation between two multivariate random variables. In our investigation, we are interested in the correlation between grammatical gender and lexical semantics for each gendered–genderless language pair. To compute this correlation, we start by solving the following optimization problem:

$$(\mathbf{a}^\star, \mathbf{b}^\star) = \underset{(\mathbf{a},\mathbf{b})}{\arg\max}\ corr(\mathbf{a}^\top G_\ell^{\text{train}}, \mathbf{b}^\top E_m^{\text{train}}).$$

Although this optimization problem is non-convex, it can be solved in closed form using singular value decomposition (SVD). We use a standard implementation of CCA (Pedregosa et al., 2011).

Having found the projections $\mathbf{a}^\star \in \mathbb{R}^{|\mathcal{G}_\ell|}$ and $\mathbf{b}^\star \in \mathbb{R}^{50}$ that maximize the correlation, we then use them to compute the correlation between grammatical gender and lexical semantics as follows:

$$\rho_{\ell,m} = corr(\mathbf{a}^{\star\top} G_\ell^{\text{test}}, \mathbf{b}^{\star\top} E_m^{\text{test}}).$$

To establish statistical significance, we follow the approach of Monteiro et al. (2016). We create $B = 100,000$ permutations of the columns of

[2] http://compling.hss.ntu.edu.sg/omw/summx.html

[3] This imbalance as a limitation of our investigation.

[4] The FASTTEXT word embeddings were trained using Common Crawl and Wikipedia data, using CBOW with position weights, with character n-grams of length 5. For more information, see http://fasttext.cc/docs/en/crawl-vectors.html.

G_ℓ^{train}; for each permutation b, we then repeat the steps above to obtain $\rho_{\ell,m}$; finally, we compute

$$p = \frac{1 + \sum_{b=1}^{B} \delta(\rho_{\ell,m}^{b} \geq \rho_{\ell,m})}{B+1}.$$

Because our investigation involves testing 90 different hypotheses, we use Bonferroni correction (Dror et al., 2017)—i.e., we multiply p by 90. If the resulting Bonferroni-corrected p-value is small, then we can reject the null hypothesis that there is no correlation between grammatical gender and lexical semantics for that language pair.

Secondarily, we investigate semantic similarities between the 18 languages' gender systems by analyzing their projections of lexical semantics. For each pair of *gendered* languages ℓ and ℓ', we compute the correlation (cosine distance) between $\mathbf{b}_\ell^\star$ and $\mathbf{b}_{\ell'}^\star$ for each of the 5 genderless languages.

4 Results

We find a significant correlation between grammatical gender and lexical semantics (i.e., the Bonferroni-corrected p-value is less than 0.05) for 55 of the 90 gendered–genderless language pairs. These results are depicted in Figure 1. For Slovak, Croatian, and Ukranian, we find no correlation for any of the genderless languages; for Slovenian, we find a significant correlation for only Mandarin Chinese. We suspect that these results are due the relatively small number of inanimate nouns considered for each of these language pairs (see Table 1 for the counts). We also find slightly different patterns of correlation for the different genderless languages

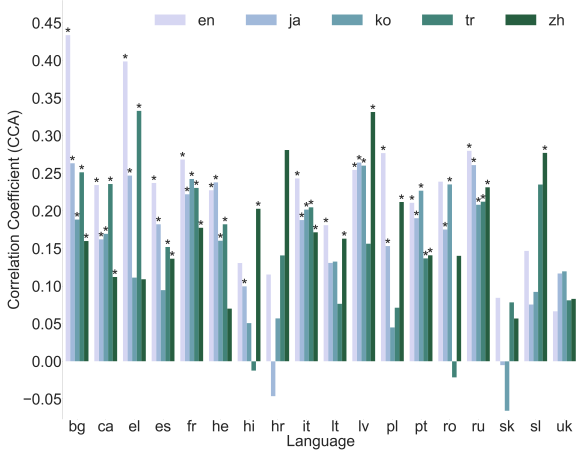

Figure 1: The correlation between grammatical gender and lexical semantics for each of the 90 gendered–genderless language pairs ($*$ indicates significance).

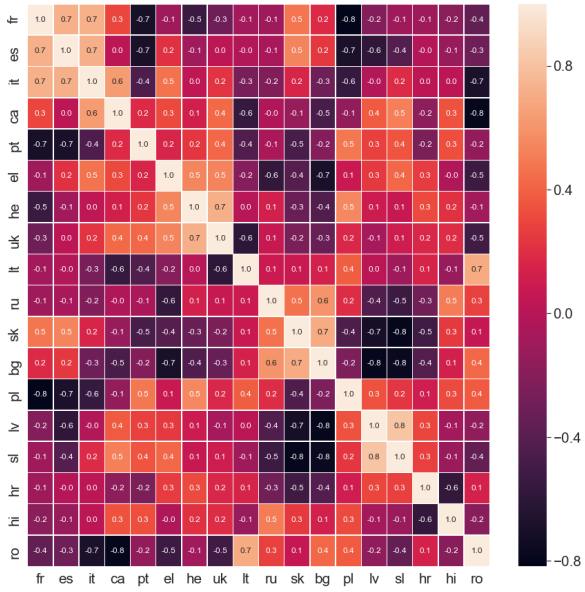

Figure 2: The correlation between $\mathbf{b}_\ell^\star$ and $\mathbf{b}_{\ell'}^\star$ for each pair of gendered languages ℓ and ℓ' (for English).

that we use to create our 5 operationalizations of lexical semantics. For Japanese, we find significant correlations for 13 of the 18 gendered languages; for English and Chinese, we find significant correlations for 12; for Korean and Turkish, we find significant correlations for 9 of the gendered languages.

For each pair of gendered languages ℓ and ℓ', Figure 2 depicts the the correlation (cosine distance) between $\mathbf{b}_\ell^\star$ and $\mathbf{b}_{\ell'}^\star$ for English. We find higher correlations for pairs of languages that are phylogenetically similar. For example, French has higher correlations with Spanish and Italian than with Polish. This is likely because phylogenetically similar languages exhibit historical similarities in their gender systems as a result of a common linguistic origin (Fodor, 1959; Ibrahim, 2014; Stump, 2015).

5 Conclusion

Our investigation is the first to quantitatively demonstrate that there is a significant correlation between grammatical gender and lexical semantics for inanimate nouns. Although our results provide evidence for the non-arbitrariness of noun–gender assignments, they must be contextualized. In contrast to animate nouns, it is not clear that a single cross-linguistic category explains our results. Moreover, we limit the scope of our investigation to frequent inanimate nouns. These nouns tend to be distributed across genders, whereas less frequent inanimate nouns tend to be assigned to a single gender (Dye et al., 2015). We leave the investigation of less frequent inanimate nouns for future work.

Acknowledgements

Thanks to Jennifer Culbertson, Jacob Eisenstein, and Arya McCarthy for conversations on this topic.

References

A.T. Aksenov. 1984. K probleme èkstralingvističeskoj motivacii grammatičeskoj kategorii roda [on extralinguistic motivation of the grammatical category of gender]. *Voprosy Jazykoznanija 33 (1)*, pages 14–25.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Proceedings of the 6th Global WordNet Conference (GWC)*, 8(4):5.

Carl Buck. 1949. *A Dictionary of Selected in the Principal Indo-European Languages*. University of Chicago Press.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Melody Dye, Petar Milin, Richard Futrell, Michael Ramscar, and Eberhard Karls. 2015. A functional theory of gender paradigms. In *Perspectives on Morphological Organization*.

John R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Istvan Fodor. 1959. The origin of grammatical gender. *Lingua*, 8:186–214.

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to Lithuanian wordnet. In *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80.

Aitor González-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Muhammad Hasan Ibrahim. 2014. *Grammatical gender: Its origin and development*, volume 166. Walter de Gruyter.

Katharina Kann and Lawrence Wolf-Sonkin. To appear. Grammatical gender, Neo-Whorfianism, and word embeddings: A data-driven approach to linguistic relativity (manuscript from 2018).

João M. Monteiro, Anil Rao, John Shawe-Taylor, Janaina Mourão-Miranda, and Alzheimer's Disease Initiative. 2016. A multiple hold-out framework for sparse partial least squares. *Journal of neuroscience methods*, 271:182–194.

Vivi Nastase and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377. Association for Computational Linguistics.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2015. Automatic expansion of Croatian wordnet. In *In Proceedings of the 29th CALS international conference Applied Linguistic Research and Methodology*, Zadar (Croatia).

Noam Ordan and Shuly Wintner. 2007. Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press.

Ida Raffaelli, Božo Bekavac, Željko Agić, and Marko Tadić. 2008. Building Croatian wordnet. In *Proceedings of the Fourth Global WordNet Conference 2008*, pages 349–359.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisaw Szpakowicz. 2012. Mapping plWordNet onto Princeton WordNet. *International Journal of Lexicography*.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Gregory Stump. 2015. Inflection classes. In *The Oxford handbook of inflection*. Oxford University Press.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American linguistics*, 21(2):121–137.

Morris Swadesh. 1971/2006. *The origin and diversification of language*. Chicago: Aldine.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452, Szeged.

Mark Twain. 1880. *A Tramp Abroad*. Wikisource.