Echoes of Persuasion: The Effect of Euphony in Persuasive Communication

Marco Guerini
Trento RISE
Povo, I-38100 Trento
marco.guerini@trentorise.eu

Gözde Özbal
FBK-Irst
Povo, I-38100 Trento
gozbalde@gmail.com

Carlo Strapparava
FBK-Irst
Povo, I-38100 Trento
strappa@fbk.eu

Abstract

While the effect of various lexical, syntactic, semantic and stylistic features have been addressed in persuasive language from a computational point of view, the persuasive effect of phonetics has received little attention. By modeling a notion of euphony and analyzing four datasets comprising persuasive and non-persuasive sentences in different domains (political speeches, movie quotes, slogans and tweets), we explore the impact of sounds on different forms of persuasiveness. We conduct a series of analyses and prediction experiments within and across datasets. Our results highlight the positive role of phonetic devices on persuasion.

1 Hocus Pocus

Historically, in human sciences, several definitions of persuasion have been proposed – see for example (Toulmin, 1958; Walton, 1996; Chaiken, 1980; Cialdini, 1993; Petty and Cacioppo, 1986). Most of them have a common core addressing: *methodologies aiming to change the mental state of the receiver by means of communication in view of a possible action to be performed by her/him.* (Perelman and Olbrechts-Tyteca, 1969; Moulin et al., 2002).

These methodologies might take into account the overall structure of a text such as the ordering of the arguments or simply single word choices. For a successful text both of them are often required. The focus of persuasion may vary according to the goal of the communication and it can take different forms according to the domain: from memorability (e.g., making people remember a statement or a product) to diffusion (e.g., making people pass on a content in social networks by sharing it), from behavioral change (e.g., political communication) to influencing purchasing decisions (e.g., slogans to convince people to try or buy a product) – see for example (Heath and Heath, 2007). While many techniques such as resorting to expert opinion, utilizing the framing effect, emotive language or exaggeration can be used to obtain such persuasive effects, we devote this study to explore particular techniques pertaining to euphony.

Euphony refers to the inherent pleasantness of the sounds of words, phrases and sentences, and it is utilized to achieve pleasant, rhythmical and harmonious effects. The idea that the pleasantness of the sounds in a sentence can foster its effectiveness is rooted in our culture, and is connected to the concepts of rhythm and music. The fact that language and music interact in our brain has been shown by localizing low-level syntactic processes of music and language in the temporal lobe (Sammler et al., 2013). It has also been shown that changes in the cardiovascular and respiratory systems can be induced by music – specifically tempo, rhythm, melodic structure (Bernardi et al., 2006). The importance of euphony has its roots also in ancient human psychology. As Julian Jaynes suggests (Jaynes, 2000), poetry used to be divine knowledge. It was the sound and tenor of authorization and it commanded where plain prose could only ask. A paradigmatic example of this conception is the act of casting a spell. Spells (incantations) are special linguistic objects that are meant not only to change how

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1483–1493,
Denver, Colorado, May 31 – June 5, 2015.

people think or behave but they are also so powerful that they can – allegedly – change reality. Spells are often very euphonic (and meaningless) sentences, e.g. "Hocus Pocus".

Various psycholinguistic studies addressed the effects of phonetics on the audience in different aspects such as memorability (Wales, 2001; Benczes, 2013) or more specifically advertisement (Leech, 1966; Bergh et al., 1984). There are also computational studies that address the problem of recognizing persuasive sentences according to various syntactic, lexical and semantic features (Danescu-Niculescu-Mizil et al., 2012; Tan et al., 2014). However, to the best of our knowledge, the direct impact of phonetic elements on persuasiveness has not been explored in computational settings yet.

In this paper, we fill in this gap by conducting a series of analyses and prediction experiments on four datasets representing different aspects of persuasive language to evaluate the importance of a set of phonetic devices (i.e. rhyme, alliteration, homogeneity and plosives) on various forms of persuasiveness. Our experiments show that phonetic features play an important role in the detection of persuasiveness and encode a notion of "melodious language" that operates both within and across datasets.

2 Related Work

In the following, we first revise some NLP studies addressing linguistic features of successful communication. Then, we summarize a selection of studies devoted to the effects of phonetics on persuasion.

2.1 NLP studies on persuasion

Berger and Milkman (2009) focus on a particular form of persuasion by using New York Times articles to examine the relationship between virality (i.e., the tendency of a content to be circulated on the Web) and emotions evoked by the content. They conduct semi-automated sentiment analysis to quantify the affectivity and emotionality of each article. Results suggest a strong relationship between affect and virality, in this case measured as the count of how many people emailed each article. As suggested by the authors, this metric represents a form of "narrowcasting", as opposed to other "broadcasting" actions such as sharing on Twitter.

Another line of research investigates the impact of various textual features on audience reactions. The work by Guerini et al. (2011) correlates several viral phenomena with the wording of a post, while Guerini et al. (2012) show that features such as the readability level of an abstract influence the number of downloads, bookmarking and citations.

A particular approach to content virality is presented by Simmons et al. (2011), who explore the impact of different types of modification on memes spreading from one person to another.

Danescu-Niculescu-Mizil et al. (2012) measure a different ingredient of persuasion by analyzing the features of a movie quote that make it "memorable". They compile a corpus consisting of memorable and non-memorable movie quote pairs and conduct a detailed analysis to investigate the lexical and syntactic differences between these pairs.

Louis and Nenkova (2013) focus on influential science articles in newspapers by considering characteristics such as readability, description vividness, use of unusual words and affective content. High quality articles (NYT articles appearing in "The Best American Science Writing" anthology) are compared against typical NYT articles.

Borghol et al. (2012) investigate how differences in textual description affect the spread of content-controlled videos. Lakkaraju et al. (2013) focus on the act of resubmissions (i.e., content that is submitted multiple times with multiple titles to multiple different communities) to understand the extent to which each factor influences the success of a content. Tan et al. (2014) consider how content spreads in an on-line community by pinpointing the effect of wording in terms of content informativeness, generality and affect. Althoff et al. (2014) develop a model that can predict the success of requests for a free pizza gifted from the Reddit community. The authors consider high-level textual features such as politeness, reciprocity, narrative and gratitude.

2.2 Studies on the effects of phonetics

Benczes (2013) states that alliteration and rhyme can be considered as attention-seeking devices as they enhance emphasis. The author also suggests that they are useful for acceptability and long-term retention of original expressions, decrypting their meanings, indicating informality, and breaking the ice be-

tween an audience and a speaker. Therefore, these devices are commonly used in original metaphorical and metonymical compounds.

According to Leech (1966), phonetic devices such as rhyme and alliteration are systematically exploited by advertisers to achieve memorability. Similarly, Wales (2001) underlines the effectiveness of alliteration and rhyme on emphasis and memorability of an expression.

The relation between the usage of plosives (i.e., consonants in which the vocal tract is blocked so that all airflow ceases, such as "p", "t" or "k") and memorability has also been investigated. According to the study carried out by Bergh et al. (1984) brand names starting with plosive sounds are recalled and recognized more than the ones starting with other sounds. Özbal et al. (2012) carry out an analysis of brand names and discover that plosives are very commonly used.

Danescu-Niculescu-Mizil et al. (2012), whom we previously mentioned, carry out an auxiliary analysis and observe the differences in letter and sound distribution (e.g. usage of labials or front vowels, back sounds, coordinating conjunctions) of memorable and non-memorable quotes.

Özbal et al. (2013) propose a phonetic scorer for creative sentence generation such that generated sentences can contain various phonetic features including alliteration, rhyme and plosive sounds. The authors evaluate the proposed model on automatic slogan generation. In a more recent work (Özbal et al., 2014), they enforce the existence of these features in the sentences that are automatically generated for second language learning to introduce hooks to echoic memory.

3 Phonetic Scorer

For the design of the phonetic features, we were mostly inspired by the work of Özbal et al. (2013), who built and used three phonetic scorers for creative sentence generation. Similarly to this work, all the phonetic features that we used are based on the phonetic representation of English words of the Carnegie Mellon University pronouncing dictionary[1]. We selected four classes of phonetic devices, namely plosives, alliteration, rhyme and homogeneity, which can easily be modeled by observing the distribution of specific classes of phonemes within the sentence. The *plosive* score is calculated as the ratio of the number of plosive sounds in a sentence to the overall number of phonemes. For both *alliteration* and *rhyme* scorers, we provide a naïve implementation that does not consider stresses or syllables, but only counts the number of repeated sounds at the beginning or end of words in the sentence. The alliteration score is calculated as the number of repeated phonetic prefixes in a sentence normalized by the total number of phonemes. Similarly, the *rhyme* score is calculated as the ratio of the number of repeated phonetic endings in a sentence to the total number of phonemes. Lastly, the homogeneity scorer simply calculates the degree of homogeneity in terms of phonemes used in a sentence independently from their positions. If we let d_{ph} be the count of distinct phonemes and t_{ph} be the total count of phonemes in a sentence, then the homogeneity score is calculated as $1 - (d_{\text{ph}}/t_{\text{ph}})$.

4 Dataset

In this section, we describe the four datasets we used to conduct our analyses and experiments. As we mentioned previously, the definition of persuasion is a debated topic and it can comprise distinct strategies or facets. For this reason, we experimented with datasets where at least one ingredient is clearly in the equation. To explore the effects of wording and euphonics on persuasion, the datasets were built in a controlled setting (topic, author, sentence length) to avoid confounding factors such as author or topic popularity, by following the procedure described in (Danescu-Niculescu-Mizil et al., 2012; Tan et al., 2014). In addition, these datasets comprise short texts (mostly single sentences) to focus on surface realization of persuasion, where strategic planning – which might act as a confounding factor – plays a minor role. The idea of using controlled experiments (usually in an A/B test setting) to study persuasive communication can be traced back at least to Hovland et al. (1953). While two of these datasets (Twitter and Movies) were already available, the other two (CORPS and Slogans) were collected by following the methodology proposed in the first two as

[1]The CMU pronunciation dictionary is freely available at `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. We have used version 0.7a in our implementation.

closely as possible[2].

All datasets are built around the core idea of collecting pairs consisting of a persuasive sentence (P) and a non-persuasive counterpart ($\neg P$), where P and $\neg P$ are structurally very similar and controlled for the above mentioned confounding factors.

Twitter. A set of 11,404 tweet pairs, where each pair comes from the same user (author control) and contains the same URL (topic control). P and $\neg P$ are determined based on their retweet counts (Tan et al., 2014). It is worth noting that in our experiments we were able to collect only 11,019 of such tweet pairs since some of them were deleted in the meanwhile.

Movie. A set of 2,198 single-sentence memorable movie quotes (P) paired with non-memorable quotes ($\neg P$). For each P, the dataset contains a contrasting quote $\neg P$ from the same movie such that (i) P and $\neg P$ are uttered by the same speaker, (ii) P and $\neg P$ have the same number of words, (iii) $\neg P$ does not occur in the IMDb list of memorable quotes and (iv) P and $\neg P$ are as close as possible to each other in the script (Danescu-Niculescu-Mizil et al., 2012).

CORPS. A set of 2,600 sentence pairs uttered by various politicians. We collected these pairs from CORPS, a freely available corpus of political speeches tagged with audience reactions (Guerini et al., 2013). The methodology that we used to build the pairs is very similar to Danescu-Niculescu-Mizil et al. (2012): for each P, where P is the sentence preceding an audience reaction (e.g. `APPLAUSE`, `LAUGHTER`), we selected a contrasting single-sentence $\neg P$ from the same speech. We required $\neg P$ to be close to P in the speech transcription, subject to the conditions that (i) P and $\neg P$ are uttered by the same speaker - which is trivial since these are monologues, where a single speaker is addressing the audience - (ii) P and $\neg P$ have the same number of words, and (iii) $\neg P$ is 5 to 15 sentences away from P. This last condition had to be imposed since, differently from movie quotes, we do not have the evidence of which fragment of the speech exactly provoked the audience reaction (i.e. it could be the combination of more than one sentence).

Slogan. A set of 1,533 slogans taken from online resources paired with non-slogans that are similar in content. We collected the non-slogans from the subset of the New York Times articles in English GigaWord – 5th Edition – released by Linguistic Data Consortium (LDC)[3]. For each slogan, we picked the most similar sentence in the New York Times articles having the same length and the highest LSA similarity (Deerwester et al., 1990) with the slogan. The LSA similarity approach that we used to collect the non-slogans is very similar to the approach used by Louis and Nenkova (2013) to collect the non-persuasive counterparts of successful news articles.

In Table 1, we sum up the criteria used in the construction of each dataset. As can be observed from the table, each dataset satisfies at least two of the three criteria described above. In the last two columns of the table, we also provide the average token length of the persuasive and non-persuasive sentences in each dataset. Finally, in Table 2 we provide examples of euphonic and persuasive sentences for each dataset together with their phonetic scores.

	Criterion			Length	
DATASET	Author	Length	Topic	P	$\neg P$
CORPS	✓	✓	✗	14.0	14.0
Movie	✓	✓	✗	9.7	9.7
Slogan	✗	✓	✓	5.0	5.0
Twitter	✓	✗	✓	16.2	15.4

Table 1: Criteria used in the construction of each dataset and average token length of persuasive and non-persuasive pairs

5 Data Analysis

To provide a first insight on the data, in Table 3 we report the average phonetic scores for each data set (Mann-Whitney U Test is used for statistical significance between P and $\neg P$ samples, with Bonferroni correction to ameliorate issues with multiple comparisons). The results are partially in line with our expectations of the euphony phenomena being more relevant in the persuasive sentences across the datasets.

As can be observed from the table, the average rhyme scores are higher in persuasive sentences and

[2]CORPS and Slogans datasets can be downloaded at the following link: `https://github.com/marcoguerini/paired_datasets_for_persuasion/`

[3]`http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T07`

Dataset	Example	Rhyme	Alliteration	Plosive	Homogeneity
CORPS	I think we can do better and I think we must do better.	0.789	0.737	0.342	0.631
	It will be waged with determination and it will be waged until we win.	0.566	0.679	0.189	0.736
Movie	The night time is the right time.	0.818	0.545	0.181	0.636
	Beautiful.... beautiful butterfly...	0.667	0.708	0.250	0.583
	Dog eat dog, brother.	0.533	0.533	0.400	0.400
Slogan	Different Stores, Different Stories.	0.621	0.896	0.207	0.690
	Why ask why? Try Bud Dry	0.625	0.625	0.312	0.437
	Live, Love, Life.	0.818	0.909	0.0	0.636
Twitter	A Nerd in Need is a Nerd indeed.	0.636	0.727	0.227	0.681
	Easter cupcake baking!!	0.0	0.0	0.412	0.470

Table 2: Euphonic examples of persuasive sentences from each dataset, along with their phonetic scores.

	Rhyme		Alliteration		Plosive		Homogeneity	
Dataset	μ	σ	μ	σ	μ	σ	μ	σ
CORPS$_{\neg P}$	0.233	0.143	0.208	0.142	0.187	0.058	0.603	0.173
CORPS$_P$	0.245†	0.152	0.223**	0.154	0.194***	0.060	0.588**	0.179
Movie$_{\neg P}$	0.196	0.143	0.167	0.142	0.191	0.073	0.485	0.155
Movie$_P$	0.214*	0.165	0.196***	0.164	0.185†	0.067	0.526***	0.164
Slogan$_{\neg P}$	0.071	0.111	0.047	0.092	0.204	0.098	0.343	0.163
Slogan$_P$	0.140***	0.194	0.123***	0.185	0.189***	0.098	0.366***	0.156
Twitter$_{\neg P}$	0.204	0.116	0.180	0.114	0.188	0.058	0.617	0.134
Twitter$_P$	0.216***	0.121	0.193***	0.120	0.185**	0.055	0.636***	0.128

Table 3: Average phonetic scores for our datasets - ***, $p < .001$; **, $p < .01$; *, $p < .05$; †, not significant

the difference is highly significant for Slogan and Twitter ($p < .001$), slightly significant for Movie quotes ($p < .05$), but not significant for CORPS. The average alliteration scores are again higher in persuasive sentences and all the differences are highly significant in all datasets (apart from CORPS with $p < .01$). Plosives seem not to correlate well with our intuition of persuasiveness and euphony: either there is no significance (movie quotes) or the averages of euphonic scores are higher in the non-persuasive sentences (the difference is highly significant in slogans, and significant in Twitter). The only dataset that meets our expectation is CORPS with a highly significant difference in favor of persuasive sentences. Finally, the average homogeneity scores are significantly ($p < .001$) higher in persuasive sentences in all datasets except CORPS, where the scores of non-persuasive sentences are significantly higher ($p < .01$) than persuasive ones.

Without going into details of cross-dataset comparisons we would like to note that CORPS seems a very peculiar dataset in terms of average scores, as compared to the others. In terms of rhyme and alliteration, the average scores of non-persuasive sentences ($\neg P$) in CORPS are always higher than the persuasive sentences (P) in the other datasets ($p < .001$ in all cases), while for homogeneity the same holds apart from Twitter. These results may derive from the fact that a political speech is a carefully crafted text – aimed at influencing the audience – in its entirety, so also "non-persuasive" sentences in CORPS are on average more persuasive than in other datasets.

As a next step, we conducted another analysis on the distribution of "extreme cases", i.e. sentences that have a very high phonetic score at least in one feature. This analysis derives from the intuition that a euphonic sentence might be recognized as such by humans only if its phonetic scores are above a certain threshold. In fact, sound repetition in a sentence may occur by chance, as in "I saw **the** knife in **the** drawer", and the longer the sentence is, the higher the probability that phonetic scores will be non-zero even in absence of a euphonic effect. Therefore, the average scores for each phonetic device, as reported in Table 3, are only partially informative.

Given this premise, to evaluate the "persuasive power" of the phonetic devices taken into account,

Dataset	$\hat{F}_{rh}(t)$	$\hat{F}_{al}(t)$	$\hat{F}_{pl}(t)$	$\hat{F}_{ho}(t)$
CORPS$_{\neg P}$	0.025	0.012	0.362	0.394
CORPS$_P$	0.033**	0.023**	0.415***	0.363†
Movie$_{\neg P}$	0.018	0.011	0.397	0.092
Movie$_P$	0.041***	0.025***	0.363†	0.173***
Slogan$_{\neg P}$	0.005	0.003	0.460	0.011
Slogan$_P$	0.055***	0.043***	0.410***	0.018***
Twitter$_{\neg P}$	0.006	0.003	0.385	0.377
Twitter$_P$	0.012***	0.008***	0.364**	0.449***

Table 4: Probability of examples above threshold, - ***, $p < .001$; **, $p < .01$; *, $p < .05$; †, not significant

we compare them in terms of empirical Complementary Cumulative Distribution Functions (CCDFs) of the persuasive/non-persuasive pairs in various datasets. These functions are commonly used to analyze online social networks in terms of growth in size and activity (see for example (Ahn et al., 2007; Jiang et al., 2010; Leskovec, 2008)) and also for measuring content diffusion, e.g. the number of retweets of a given content (Kwak et al., 2010). Here, we use CCDFs to account for the probability P that the score of a phonetic device d will be greater than n indicating it with $\hat{F}_d(n)$. For example, the probability of having a text with more than .75 rhyme score is indicated with $\hat{F}_{rh}(.75) = \mathrm{P}(\#\mathrm{rhyme} > .75)$. To assess whether the CCDFs of the several types of texts we take into account show significant differences, we use the Kolmogorov-Smirnov goodness-of-fit test, which specifically targets cumulative distribution functions. In particular, for each phonetic device and dataset, we use a two-tailed Kolmogorov-Smirnov test (again with Bonferroni correction) to test whether the number of examples above the threshold is higher in the persuasive sentences than in their non-persuasive counterparts for that device.

Since we do not have a theoretical way to define such thresholds, we resort to empirically define them by using a specific dataset of euphonic sentences. Even if it might seem reasonable to consider poems as paradigmatic examples of "euphonic" writing, we discard them as the phonetic devices used in poems may span across sentences. Instead, we resort to tongue twisters as a gold reference of how a euphonic sentence should be. Accordingly, we collected a set of 534 tongue twisters from various online resources. Then, for each phonetic index we defined our thresholds as the average of the phonetic scores in this data, in particular: $t_{rh} = 0.55$ for rhyme, $t_{al} = 0.58$ for alliteration, $t_{pl} = 0.20$ for plosives and $t_{ho} = 0.68$ for homogeneity.

In Table 4, we report the results of our CCDF analysis. After analyzing the "extreme cases", where euphony is granted, we see that the trends found in Table 3 on the correlation between persuasiveness and euphony are confirmed and strengthened. The number of persuasive sentences with a rhyme score above threshold is 30% more than the non-persuasive ones in CORPS, while the difference is 90% in Twitter[4]. The ratio of persuasive sentences above threshold to non-persuasive ones is very high in movies and slogans (more than 2 and 10 respectively). All results are either highly significant or significant. For comparison, in Table 3 these differences are not significant for CORPS and only slightly significant ($p < .05$) for movies. Concerning alliteration, there are 85% more cases above threshold in the persuasive sentences of CORPS than the non-persuasive ones. For movie quotes and Twitter, the persuasive sentences above threshold are more than two times as many as the non-persuasive ones, while the ratio is more than 13 for slogans. All results are either highly significant or significant in line with the results of Table 3. Instead, for plosive scores we observe a negative or no correlation with persuasiveness, the only exception being CORPS. Regarding homogeneity, for CORPS the difference between persuasive and non-persuasive sentences is not significant (in Table 3 it was significantly in favor of non-persuasive sentences), while for the other datasets there is a highly significant difference in favor of persuasive sentences (between 20% and 80%). As a whole, these results confirm our intuition that phonetic features play a significant role with respect to persuasiveness. In the next section we will validate this claim by means of prediction experiments.

6 Prediction Experiments

In this section, we describe the prediction tasks (both within and across datasets) that we carried out to in-

[4]In the following the ratios are computed on the real values while Table 4 presents the rounded values.

vestigate the impact of the phonetic features on the detection of various forms of persuasiveness. We compare three different sets of features, namely phonetic, n-grams and their combination to understand whether phonetic information can improve the performance of standard lexical approaches. Similarly to Danescu-Niculescu-Mizil et al. (2012) and Tan et al. (2014), we formulate a pairwise classification problem such that given a pair (s_1, s_2) consisting of sentences s_1 and s_2, the goal is to determine the more persuasive one (i.e., the one on the *left* or *right*). We can consider this as a binary classification task where for each instance (i.e., pair) the possible labels are *left* or *right*.

6.1 Dataset and preprocessing

For the prediction experiments, we used the four datasets described in Section 4 (i.e., CORPS, Twitter, Slogan and Movie), all of which consist of a persuasive sentence P and its non-persuasive counterpart ($\neg P$) labeled as either *left* or *right*. To make the positions of the sentences in a pair irrelevant (i.e. to provide symmetry), for each instance occurring in the original datasets (e.g., (s_1, s_2) with label *left*), we added another instance including the same sentence pair in reverse order (i.e., (s_2, s_1) with label *right*). As a preprocessing step, all the sentences were tokenized by using Stanford CoreNLP (Manning et al., 2014).

6.2 Classifier and features

We performed a 10-fold cross-validation on each dataset and experimented with three feature sets by using a Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995). We preferred SVM as a classifier due to its characteristic property to especially perform well on high-dimensional data (Weichselbraun et al., 2011).

The first feature set consists of the phonetic features (i.e. plosive, alliteration, rhyme and homogeneity scores as detailed in Section 3). The second feature set is a standard bag of word n-grams including unigrams, bigrams and trigrams. All the non-ascii characters, punctuations and numbers were ignored. The URLs and mentions in Twitter data were replaced with tags (i.e. *_URL_* and *_MENTION_* respectively). In addition, for the unigram features, stop words were filtered out. We did not apply this filtering for bigrams and trigrams to capture longer-range usage patterns such as propositional phrases. The third feature set is simply the union of both phonetic and n-gram features.

To find the best configuration for each dataset and feature set, we conducted a grid search over the degree of the polynomial kernel (1 or 2) and the number of features to be used (in the range between 1,000 and 20,000). Due to the low dimensionality of the phonetic feature set, feature selection was performed only for the feature sets including n-grams. The selection was performed based on the information gain of each feature.

Dataset	*Phonetic*	*N-Gram*	*All*
CORPS	0.589 (-, 1)	0.733^{***} (4k, 1)	$0.736^{\dagger}$ (2k, 1)
Movie	0.600 (-, 2)	0.694^{***} (1k, 1)	0.722^{***} (1k, 1)
Slogan	0.700 (-, 2)	0.826^{***} (3k, 1)	0.883^{***} (5k, 1)
Twitter	0.563 (-, 2)	0.732^{***} (5k, 1)	0.745^{***} (4k, 1)

Table 5: Results of the within-dataset experiments.

6.3 Within-dataset experiments

For this set of experiments, we conducted a 10-fold cross validation on each dataset separately. In Table 5, for each dataset listed in the first column, in the subsequent columns we report the performance of the best model obtained with 10-fold cross validation using i) only phonetic features (*Phonetic*), ii) only n-grams (*N-Gram*), iii) both phonetic and n-gram features (*All*). As mentioned previously, for each pair (s_1, s_2) consisting of sentences s_1 and s_2, our dataset contains another pair including the same sentences in reverse order (i.e., (s_2, s_1)), resulting in a symmetric and balanced dataset. Therefore, classification performance is measured in terms of accuracy (i.e., the percentage of pairs of which labels were correctly predicted). For each accuracy value, we also report in parenthesis the number of features selected and the kernel degree of the corresponding model. While the kernel degree did not make a big difference in the performance, the number of selected features had an important effect on the accuracy of the models. As can be observed from these values, the best performance on all the datasets is achieved with a relatively small number of features.

Among the values reported in the table, the ones followed by *** are significantly different ($p < .001$)

Dataset	*N-Gram*	*N-Gram+Rhyme*	*N-Gram+Plosive*	*N-Gram+Homogeneity*	*N-Gram+Alliteration*
CORPS	0.733	$0.738^{\dagger}$ (3k, 1)	$0.740^{\dagger}$ (2k, 1)	$0.738^{\dagger}$ (3k, 1)	$0.738^{\dagger}$ (2k, 1)
Movie	0.694	$0.694^{\dagger}$ (1k, 1)	$0.692^{\dagger}$ (1k, 1)	0.721^{***} (1k, 1)	0.709^{**} (1k, 1)
Slogan	0.826	0.864^{***} (2k, 1)	$0.824^{\dagger}$ (2k, 1)	0.867^{***} (3k, 1)	0.859^{***} (3k, 1)
Twitter	0.732	0.740^{**} (4k, 1)	$0.733^{\dagger}$ (4k, 1)	0.746^{***} (4k, 1)	0.742^{***} (4k, 1)

Table 6: Contribution of the phonetic features.

	Test											
	CORPS			Twitter			Slogan			Movie		
Training	*Phonetic*	*N-Gram*	*All*	*Phonetic*	*N-Gram*	*All*	*Phonetic*	*N-Gram*	*All*	*Phonetic*	*N-Gram*	*All*
CORPS	-	-	-	0.463	0.508	0.523	0.508	0.517	0.539	0.411	0.506	0.516
Twitter	0.439	0.494	0.462	-	-	-	0.564	0.531	0.637	0.596	0.544	0.589
Slogan	0.535	0.512	0.514	0.535	0.510	0.539	-	-	-	0.532	0.545	0.588
Movie	0.431	0.513	0.498	0.562	0.533	0.560	0.581	0.537	0.589	-	-	-

Table 7: Results of the cross-dataset prediction experiments optimized on the training set.

from the ones to their left, while † represents no significance, as calculated according to McNemar's test (McNemar, 1947). For each dataset, the weakest models (i.e. the ones using only the phonetic features in all cases) are still significantly ($p < .001$) more accurate than a random baseline (accuracy = 50%). As can be observed from the table, the models using only n-grams significantly outperform the ones only based on phonetic features in all datasets. However, while the phonetic features are not very strong by themselves, their combination with n-grams results in models outperforming the n-gram based models in all cases. The difference is highly significant for all datasets except CORPS, where n-grams alone are sufficient to achieve a good performance. We speculate that the kind of persuasiveness used in political speeches is more dependent on the lexical choices of the speaker and on the use of a specific set of semantically loaded words such as *bless*, *victory*, *God* and *justice* or *military*. This is in line with the work of Guerini et al. (2008), who built a domain specific lexicon to study the persuasive impact of words in political speeches.

We also conducted an additional set of experiments to investigate if some phonetic features stand out among the others, and to find out the contribution and importance of each phonetic feature in isolation. To achieve that, for each dataset we conducted a 10-fold cross validation to obtain the best four models containing a single phonetic feature on top of n-gram features (i.e. *N-Gram+Rhyme*, *N-Gram+Plosive*, *N-Gram+Homogeneity* and *N-Gram+Alliteration*). In Table 6, we report the accuracy of the n-gram model and these four models for each dataset. Similarly to Table 5, for each accuracy value, we also report in parenthesis the number of features selected and the kernel degree of the corresponding model obtained with grid search. The results demonstrate that homegeneity is the most effective feature when added on top of n-grams, resulting in highly significant improvement against the basic n-gram models in three out of four datasets. Alliteration and rhyme closely follow homogeneity by yielding models that significantly outperform the n-gram models in three and two datasets respectively. Finally, the models containing plosives do not improve over the n-gram models in any of the four datasets. It is worth noting that in CORPS none of the n-gram models enriched with phonetic features improves over the basic n-gram models as in line with the results of the within-dataset experiments reported in Table 5.

6.4 Cross-dataset experiments

After observing that the combination of phonetic and n-gram features can be effective in the within-dataset prediction experiments, we took a further step and investigated the interaction of the three feature sets across datasets. More specifically, we classified each dataset with the best models (one for each feature set) trained on the other datasets. With these experiments, we investigated the ability of phonetic

features to generalize across the different lexicons of the datasets. As we discussed previously, the four datasets represent different forms of persuasiveness. In this respect, the results of the cross-dataset experiments can also be interpreted as a measure of the degree of compatibility among these kinds of persuasiveness.

In Table 7, we present the results of the cross-dataset prediction experiments. For each training and test set pair, we report the accuracy of the best models, one for each feature set, based on cross-validation on the training set. As can be observed from the table, the figures are generally low and various domain adaptation techniques could be employed to improve the results. However, the objective of this evaluation is not to train an optimized cross-domain classifier, but to assess the potential of the feature sets to model different kinds of persuasiveness.

As expected, n-gram features show poor performance due to the lexical and stylistic differences among the datasets. In many cases, the phonetic models outperform the n-gram models, and in several cases the combination of the two feature sets deteriorates the performance of the phonetic features alone. These findings support our hypothesis that phonetic features, due to their generality, have better correlation with different forms of persuasiveness than lexical features. The experiments involving the CORPS dataset, both for training and testing, do not share this behavior. Indeed, when CORPS is used as a training or test dataset, the performance of the models is quite low (very close to or worse than the baseline in many cases) independently from the feature sets. These results suggest that the notion of persuasiveness encoded in this dataset is remarkably different from the others, as previously discussed in the data analysis in Section 5. As seen in the within dataset experiments (see Table 5), CORPS is the only dataset in which the combination of lexical and phonetic features do not improve the classification accuracy. This explains the inability of the phonetic features to improve the accuracy in cross-dataset experiments when this dataset is employed.

7 Conclusion

In this paper, we focused on the impact of a set of phonetic features – namely rhyme, alliteration, homogeneity and plosives – on various forms of persuasiveness including memorability of slogans and movie quotes, re-tweet counts of tweets, and effectiveness of political speeches. We conducted our analysis and experiments on four datasets comprising pairs of a persuasive sentence and a non-persuasive counterpart.

Our data analysis shows that persuasive sentences are generally euphonic. This finding is confirmed by the prediction experiments, in which we observed that phonetic features consistently help in the detection of persuasiveness. When combined with lexical features, they help improving classification performance on three of the four datasets that we considered. The key role played by phonetic features is further underlined by the cross-dataset experiments, in which we observed that phonetic features alone generally outperform the lexical features. To the best of our knowledge, this is the first systematic analysis of the impact of phonetic features on several types of persuasiveness. Our results should encourage researchers dealing with different aspects of persuasiveness to consider the inclusion of phonetic attributes in their models.

As future work, we will investigate the impact of other phonetic devices such as assonance, consonance and rhythm on persuasiveness. It would also be interesting to focus on the connection between sound symbolism and persuasiveness, and investigate how the context or domain of persuasive statements interacts with the sounds in those statements.

We would like to conclude this paper with the most favorite and retweeted tweet of @NAACL2015 (the Twitter account of the conference whose proceedings comprise this paper), which is a good example of the positive effect of euphony in persuasiveness:

The deadline for @NAACL2015 paper submissions is approaching: Remember, remember, the 4th of December!

Acknowledgments

This work has been partially supported by the Trento RISE PerTe project.

References

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pages 835–844. ACM.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. *Proocedings of ICWSM*.

Rka Benczes. 2013. The role of alliteration and rhyme in novel metaphorical and metonymical compounds. *Metaphor and Symbol*, 28(3):167–184.

Jonah A. Berger and Katherine L. Milkman. 2009. Social Transmission, Emotion, and the Virality of Online Content. *Social Science Research Network Working Paper Series*, December.

Bruce G. Vanden Bergh, Janay Collins, Myrna Schultz, and Keith Adler. 1984. Sound advice on brand names. *Journalism Quarterly*, 61(4):835, dec.

Luciano Bernardi, Cesare Porta, and Peter Sleight. 2006. Cardiovascular, cerebrovascular, and respiratory changes induced by different types of music in musicians and non-musicians: the importance of silence. *Heart*, 92(4):445–452.

Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. 2012. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1186–1194. ACM.

Shelly Chaiken. 1980. Heuristic vs. sistematic information processing and the use of source vs message cues in persuasion. *Journal of Personality and Social Psychology*, 39:752–766.

Robert B. Cialdini. 1993. *Influence. The psychology of persuasion*. William Morrow & Company, Inc., New York.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the ACL*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Corps: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, 5(1):19–32.

Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *Proceedings of ICWSM-11*, Barcelona, Spain, July.

Marco Guerini, Alberto Pepe, and Bruno Lepri. 2012. Do linguistic style and readability of scientific abstracts affect their virality. *Proceedings of ICWSM-12*.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The new release of corps: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.

Chip Heath and Dan Heath. 2007. *Made to stick: Why some ideas survive and others die*. Random House.

Julian Jaynes. 2000. *The origin of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin Harcourt.

Jing Jiang, Christo Wilson, Xiao Wang, Peng Huang, Wenpeng Sha, Yafei Dai, and Ben Y Zhao. 2010. Understanding latent interactions in online social networks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 369–382. ACM.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *ICWSM*.

Geoffrey N. Leech. 1966. *English in advertising : a linguistic study of advertising in Great Britain / [by] Geoffrey N. Leech*. Longmans London.

Jurij Leskovec. 2008. *Dynamics of large networks*. ProQuest.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *TACL*, 1:341–352.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, jun.

Bernard Moulin, Hengameh Irandoust, Micheline Belanger, and Gaëlle Desordes. 2002. Explanation and

argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, 17:169–222.

Gözde Özbal, Carlo Strapparava, and Marco Guerini. 2012. Brand pitt: A corpus to explore the art of naming. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Gözde Özbal, Daniele Pighin, and Carlo Strapparava. 2013. BRAINSUP: Brainstorming Support for Creative Sentence Generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1446–1455, Sofia, Bulgaria, August. Association for Computational Linguistics.

Gözde Özbal, Daniele Pighin, and Carlo Strapparava. 2014. Automation and evaluation of the keyword method for second language learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357. Association for Computational Linguistics.

Chaim Perelman and Lucie Olbrechts-Tyteca. 1969. *The new Rhetoric: a treatise on Argumentation*. Notre Dame Press.

Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19:123–205.

Daniela Sammler, Stefan Koelsch, Tonio Ball, Armin Brandt, Maren Grigutsch, Hans-Jürgen Huppertz, Thomas R Knösche, Jörg Wellmer, Guido Widman, Christian E Elger, et al. 2013. Co-localizing linguistic and musical syntax with intracranial eeg. *NeuroImage*, 64:134–146.

Matthew Simmons, Lada A Adamic, and Eytan Adar. 2011. Memes online: Extracted, subtracted, injected, and recollected. *Proceedings of ICWSM-11*.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *ACL*.

Stephen Toulmin. 1958. *The Use of Arguments*. Cambridge University Press, Cambridge MA.

Katie Wales. 2001. *A dictionary of stylistics*. Harlow, Eng. ; New York : Longman, 2nd ed edition. Includes bibliographical references (p. 413-429).

Douglas N. Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Albert Weichselbraun, Stefan Gindl, and Arno Scharl. 2011. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1053–1060, New York, NY, USA. ACM.