Understanding the Internet: Psychological word norms as indicators of query-specific internet word frequencies

Ludwig Maximilians-Universität München

Jörg-Henrik Heine & Matthias Spörrle

Abstract

By using existing psychological word norms obtained by rating procedures we try to predict the frequencies of search hits derived from internet search engines. We used several major search engines and repeated measurement to develop a highly reliable scale of internet word frequency. We presumed that psychological criteria like typicality and valence of nouns predict the frequencies of search-operator-specific frequencies of internet search hits. Regression analysis confirmed this assumption indicating that the verbal content of the internet as interactive mass medium can be predicted by already existing and established psychological characteristics of words.

Introduction

Since the user spectrum of the internet highly has expanded in recent years it seems reasonable to assume that the verbal content of the internet is representative for usual (written) verbalizations. In summary, research indicates that the internet is used for communication purposes to a large extent (Kraut, Patterson, Lundmark, Kiesler, Mukopadhyay, & Scherlis, 1998), that the content of this communication is comparable with that of offline communication (Stern, 2004) and that contributors of internet content do not seem to substantially differ from non-contributors regarding their personality (Marcus, Machilek, & Schütz, 2006).

In line with the assumption that the internet can be seen as a large world wide multi-author text corpus, earlier research confirms a substantial relation between linguistic text corpora and the written internet content in regard of word frequency (Keller & Lapata, 2003; Blair, Urland, & Ma 2002). Consistent with the assumption that typical exponents of a category will occur more frequently when producing a piece of text Hernández-Munoz, Izura, and Ellis (2006) found a significant correlation of r = .29 between typicality and frequency. Moreover there is evidence that computational linguistic models of similarity and co-occurrence are substantial correlated to human ratings of similarity (McDonald, 1997).

In the present exploratory study we performed several internet queries using compound search terms build out of two words, one describing the category and another being a more or less typical exponent of the given category respectively. As a least equivalent to computational linguistic models of co-occurrence both operands were either put together in one quote or were connected using the Boolean [AND]. We presumed that existing psychological word norms like typicality and valence of nouns predict the frequencies of the search-operator-specific internet search hits.

Method

Four search engines were selected for the internet query: Abacho, Google, MSN, and Yahoo. Serving as a control measure, the linguistic text corpus „The German Reference Corpus DeReKo“, provided by the Institute for German Language, was also queried, using the WWW-application COSMAS II (Corpus Search, Management and Analysis System), Version 1.1. The test sample consisted of 95 commonly used German terms of profession taken from Spörrle and Rudolph (2000). Three different ways of querying the internet were used:

1) Single word search corresponding to typing in one word (i.e., occupation) into the search engine.
2) Conjunction search corresponding to typing in two words (i.e., the specific job title and the label of the category [occupation/Beruf] connected with a Boolean [AND] into the search engine.
3) Phrases search corresponding to typing in these two words in quotes into the search engine.

The internet query was done automatically by using a customized desktop script. For five times of measurement within September and October in 2006 all 95 terms of profession were sampled using the selected internet search engines and COSMAS II.

Table 1.

Ranked Pearson Intercorrelation between the rank transformed Internet scale, COSMAS II and the psychological word norms from Spörrle and Rudolph (2000)

Ranked Pearson	1	2	3	4	5	6	7	8	9	10	11	12
						Professions (n = 95)						
1. Prestige	-	.87	-.03	.90	-.83	-.68	.29	.41	.38	.37	.46	.46
2. Sympathy		-	.01	.93	-.93	-.78	.29	.37	.41	.43	.51	.47
3. Typicality			-	.17	-.10	.03	.07	.17	.34	.14	.24	.34
4. Positivity				-	-.91	-.78	.28	.41	.47	.43	.53	.53
5. Negativity					-	.82	-.26	-.36	-.39	-.37	-.46	-.42
6. Ambivalence (Griffin)						-	-.28	-.37	-.40	-.39	-.44	-.39
7. Internet scale single word							-	.90	.59	.74	.70	.56
8. Internet scale [AND]								-	.73	.75	.79	.67
9. Internet scale phrases									-	.64	.75	.81
10. COSMAS II single word										-	.95	.74
11. COSMAS II [AND]											-	.81
12. COSMAS II phrases												-

Note: Correlations stronger than | r | > .18 are significant p < .05, correlations stronger than | r | > .28 are significant p < .01

Results

Absolute frequency results returned from the search engines were ranked and the following statistical analyses were performed with the ranked data, as described by Conover and Iman (1981).

Reliability

Results from the corpus were constant over five days of measurement. The retest reliability for each search engine between the five days of measurement, based on the rank transformed raw data, ranged from r_{tt} = 0.93 to r_{tt} = 1.00 for each of the three ways of querying the internet and all correlations were significant p < .01. Furthermore, a reliability analysis for the frequency results using the different search engines was calculated using the rank transformed medians out of the results over the five days of measurement of the four internet search engines. Cronbach's Alpha for these scales ranged between r = .94 and r = .95 in this analysis, depending on three ways of querying the internet.

Correlation

Correlations between the psychological word norms from Spörrle and Rudolph (2000), the internet frequency scales from the four search engines and the frequency results out of COSMAS II were calculated. A change in strength of correlation between typicality and internet frequencies using different search strategies can bee seen (see Table 1.).

Regression

Hierarchical regression analysis (see Table 2) confirmed the assumption that existing psychological variables can predict the obtained internet frequency results. Depending on the used search strategy sympathy and prestige have proved to be valid predictors of the internet frequency scale in the first steps of the regressions. In the second steps, the incremental predictive value of typicality systematically increased as the search strategy increased its association with the superordinate category: When using phrases search, a significant increment in explained variance was found .

Table 2.

Summary of hierarchical regression analysis for rank transformed word norms (Spörrle & Rudolph 2000) predicting rank transformed internetscales using different search terms (N=95)

	Single Word Search		
	B	*SE B*	β
Step 1			
Sympathy	.15	.20	.15
Prestige	.16	.20	.16
Step 2			
Sympathy	.14	.20	.14
Prestige	.17	.20	.17
Typicality	.07	.10	.07

Note: R^2_{adj} = .07 for Step 1 (p < .05); R^2_{adj} = .07, ΔR^2 = .01 for Step 2 (p > .05).

	Boolean Search		
	B	*SE B*	β
Step 1			
Sympathy	.06	.19	.06
Prestige	.36	.19	.36+
Step 2			
Sympathy	.03	.19	.03
Prestige	.39	.19	.39*
Typicality	.18	.09	.18+

Note: R^2_{adj} = .15 for Step 1 (p < .001); R^2_{adj} = .17, ΔR^2 = .03 for Step 2 (p < .10). *p < .05; +p < .10

	Phrases Search		
	B	*SE B*	β
Step 1			
Sympathy	.31	.19	.31
Prestige	.12	.19	.12
Step 2			
Sympathy	.25	.18	.25
Prestige	.17	.18	.17
Typicality	.34	.09	.34*

Note: R^2_{adj} = .15 for Step 1 (p < .001); R^2_{adj} = .26, ΔR^2 = .12 for Step 2 (p < .001). *p < .05

Discussion

- Results of the present study confirmed the findings of earlier research indicating a relation between web and corpus word frequencies.
- Consistent with the assumptions concerning co-occurrence to reflect semantic similarity (McDonald, 1997), results of the present study support the conclusion that the psychological variable typicality can be operationalized appropriately in a way combining the word describing the category and the words describing its exponents as a compound search term put in quotes.
- For future psychological research word norms applying to typicality may be more easily gathered by querying the internet and thus are available specifically for the question of research in the moment when needed.

References

Blair, I.V., Urland, G.R., & Ma, J.E. (2002). Using internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, and Computers, 34*(2), 286-290.

Conover, W.J., & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Hernández-Munoz, N., Izura, C., & Ellis A. W. (2006). Cognitive aspects of lexical availability. *European Journal of Cognitive Psychology, 18*(5), 730-755.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics, 29*(3), 459-484.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., & Scherlis, W. (1998). Internet paradox, a social technology that reduces social involvement and psychological well-being? *American Psychologist, 53*(9), 1017-1031.

Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology, 90*(6), 1014-1031.

McDonald, S. (1997, June). *Exploring the validity of corpus-derived measures of semantic similarity.* Paper presented at the 9th Annual CCS/HCRC Postgraduate Conference, University of Edinburgh.

Spörrle, M., & Rudolph, U. (2000). Was machen Sie beruflich? Evaluative Einstellungen und Wortnormen für Berufsbezeichnungen im Deutschen [What is your occupation? Evaluative attitudes and word norms for terms of occupation in German]. *Zeitschrift für Experimentelle Psychologie, 47*(4), 297-307.

Stern, S. R. (2004). Expressions of identity online: Prominent features and gender differences in adolescents' world wide web home pages. *Journal of Broadcasting and Electronic Media, 48*(2), 218-243.

Contact: jhheine@googlemail.com