Running head: INDIVIDUALISM AND PRONOUN USE

Individualism and First Person Pronoun Use in Written Texts Across Languages*

Irem Uz

TOBB University of Economics and Technology

Department of Psychology

iuz@etu.edu.tr

Version: 08/10/2014

Abstract

The previous studies showed that the use of first person singular pronoun, i.e., "I", primes the independent self, whereas the use of the first person plural pronoun, i.e., "We", primes the interdependent self. In a different line of research, Kashima and Kashima (2003) found a strong correlation between societal level of individualism and the overall requirement of explicit personal pronoun use in a language. The present study provided a competitive test of these two hypotheses by utilizing Google Books ngram database of published books in different languages. The sample consisted of published work in American English, British English, English publications other than American and British English, Chinese, French, German, Hebrew, Italian, Russian, and Spanish corpora. The results extended the previous individual level priming experiments on first person singular vs. plural pronouns, and showed that the relative percentage of first person singular pronoun use in written material across languages reflected the cultural level individualism. However, there was not a reliable relation between the overall use of explicit pronouns and societal level of individualism.

(172 words)

KEYWORDS: Individualism

Personal Pronoun

Language

Cultural Psychology

Cultural Differences

Values, Attitudes, Beliefs

Individualism and First Person Pronoun Use in Written Texts Across Languages

Language and cultural values are inextricably linked. Language influences culture (e.g., Brewer & Gardner, 1996) and culture influences language (e.g., Na & Choi, 2009). The present paper focuses on a grammatical feature of language and individualism. Specifically, it is suggested that the rate at which particular personal pronouns are used in a language and levels of individualism are related.

The strongest claim about the relation between language and culture was made by Sapir (1912) and Whorf (1956) who proposed that patterns of thought are constrained by linguistic structure. Later researchers adopted the weak form of this hypothesis as language influencing cognition (e.g., Lau et al., 2004; Slobin, 1996). Consistent with this weak form of Sapir-Whorf hypothesis, psychologists have documented that language does influence responses of research participants. Languages whose host culture values individualism biases responses toward individualism, and languages whose host culture values collectivism biases responses toward collectivism (e.g., Ayyash-Abdo, 2001; Kemmelmeier & Cheng, 2004; Ross, Xun & Wilson, 2002; Trafimow, Silverman, Fan & Law, 1997; Watkins & Gerong, 1999).

Aside from such a general language effect on culture, structural features of languages have also been linked to levels of individualism. Kashima and Kashima (1998; 2003) focused on the relation between personal pronouns (i.e., I, you, s/he, we, and they) and individualism. Some languages require explicit use of personal pronouns, whereas others allow for drop of pronouns (Haspelmath, Dryer, Gil, & Comrie, 2005). Kashima and Kashima (1998) proposed that explicit use of pronouns in a language brings the actor to the foreground, whereas drop of pronouns deemphasizes actor and necessitates reliance on contextual cues. It is known that focus on actors rather than context is a characteristic of individualism (e.g., Choi, Nisbett & Norenzayan, 1999; Morris & Peng, 1994; Oyserman, Coon & Kemmelmeier, 2002). Kashima

and Kashima (1998; 2003) found that there was a strong positive correlation between pronoun requirement in a language and average levels of individualism across countries speaking that language even after controlling for income and latitude.

In a different line of research, several researchers have found that the use of first person singular pronoun ("I") within a particular language primes the independent or individualist self, whereas the use of the first person plural pronoun ("We") primes the interdependent or collectivist self (e.g., Brewer & Gardner, 1996; Gardner, Gabriel, & Lee, 1999; van Baaren, Maddux, Chartrand, de Bouter, & van Knippenberg, 2003). Kashima and Kashima (1998; 2003) did not attempt to account for the effect of the use of first person singular pronouns over first person plural pronouns. Instead, they focused on the effect of overall presence of personal pronouns per se. However, if using the first person singular pronouns ("I") is mandatory in a language, so are the first person plural ("we") and second ("you") or third ("he, she, it") person pronouns for that matter (Haspelmath, Dryer, Gil, & Comrie, 2005). When personal pronouns are explicitly used, the particular pronoun selected among many depends on the cultural orientations: people with chronic or temporary individualist selves prefer first person plural pronouns over first-person singular pronouns (Na & Choi, 2009). Likewise, the relative use of first person singular vs. plural pronouns is related to levels of individualism: changes in the use of first person singular vs. plural pronouns over the years within written products in US were related to the rise of individualism (Twenge, Campbell, & Gentile, 2013). Given the findings that the use of "I" vs. "we" is differentially related to individualism, a language requiring the use of personal pronouns would not only highlight the individualist self when first person singular pronoun is used, but also it would highlight the collectivist self when first person plural pronoun is used.

In summary, there are two lines of research which seem to challenge each other. One is the pronoun study (Kashima & Kashima, 1998; 2003) showing that overall pronoun use is

correlated with individualism. The other is experimental research showing that it is the type of the pronouns that matters for levels of individualism. On surface, it is possible but not likely that these two are independent effects not opposing each other: to the extent that the use of "I" is more frequent than the use of "we" across no-prodrop languages and the relative frequency of "I" vs. "we" is greater in no-prodrop languages compared to prodrop languages, one can find that in languages requiring explicit pronouns, people may came across with "I" more frequently, and thus, be more individualist, ceteris paribus. Should this be not possible, then, to the extent that the effect of overall pronoun use is stronger than the effect of relative use of "I" over "we", one can find peoples of no-prodrop languages to be more individualist than the peoples of prodrop languages.

The goal of the present study was to simultaneously test, for the first time, the relative pronoun use hypothesis and the overall pronoun use hypothesis. It aimed to extend findings pertaining to the relative pronoun use in a single language to a multi-language sample, and to test the overall pronoun use hypothesis (Kashima & Kashima, 1998; 2003) by using continuous measures of the actual frequency of pronoun use across languages rather than the dichotomous drop vs. no-drop categorization. In doing so, Google Books ngram database was used. Google Books ngram is an extensive database containing about 4-6% of all books ever published (Lin, Michel, Aiden, Orwant, Brockmann, & Petrov, 2012; Michel et al., 2011a). Twenge et al. (2013) successfully used Google Books ngram database in American English to analyze the change in pronoun use in English over the years. Following in their footsteps, the present study focuses on corpora available in American English, British English, all English, Chinese, French, German, Hebrew, Italian, Russian and Spanish. Utilizing ngram database, it was possible to test whether the relative use of first person singular over plural pronouns and overall use of pronouns were related to national levels of individualism.

Method

Overview

The relative frequency of first person singular pronoun and first person plural pronouns as well as the overall frequency of pronouns were retrieved from Google ngram database. Individualism scores were taken from Hofstede, Hofstede, and Minkov (2010). The unit of analysis was language. The total number of languages and variants were 10.

Google ngram Database

Google ngram utilizes Google Books. Google Books is a project to digitize all the material ever published. Google partners with more than 100 libraries around the world and individual publishers for its supply of books. Google ngram is based upon about 4-6% of all books ever printed (Lin et al., 2012; Michel et al., 2011a). They were composed of both fiction and non-fiction material[1] but excludes periodicals such as journals, almanacs etc. (Michel et al., 2011b). As Michel et al. (2011a) noted: "The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years". With the Google ngram, it is possible to get the frequency of a word across the vast corpora within seconds.

In its current state, Google ngram is database of languages, not countries except for US and UK. Country of publication data are recoded only for US and UK. Google Books ngram maintain database in American English, British English, all English, Chinese, French, German, Hebrew, Italian, Russian and Spanish, making it especially suitable for language based analyses.

---

[1] No distinction was maintained between fiction and non-fiction in Google ngram database of languages except for a seperate "all English fiction" corpus. However, amount of fiction vs. non-fiction was shown not to matter for analysis using pronouns as variables (e.g. Twenge et al., 2013). Further, the percentage of literature books within total number of books published in a country (UNESCO, p.39-40, 1982), and the level of individualism (Hofstede, Hofstede, & Minkov, 2010) are not correlated, $r = .041$, $p = .789$.

Ngrams in Google are constructed by counting number of single words (for 1-grams) or a series of words (for 2-5-grams) in published material in a language. For each language, there is a time series. Across the time series, the data are standardized by the total number of words available for a year (Michel et al., 2011b). Therefore, it is possible to compare the data across time as well as across languages. The reliability of a corpus increases as the proportion of books scanned increases. The total number of books included in the construction of ngrams corpus in 2009 is 4,541,627 for books in English, 854,649 for books in Spanish, 792,118 for books in French, 657,991 for books in German, 591,310 for books in Russian, 305,763 for books in Italian, 302,652 for books in Chinese, and 70,636 for books in Hebrew (Lin et al., 2012). Of the countries speaking those languages, US, China, Germany, France, UK, Italy, Spain, Canada, Australia, Russia, Belgium, and Switzerland alone accounts for 86% of the world publishing market (International Publishers Association, 2013; - excluding Japan, Brazil, Korea, India, Turkey, Netherlands, Poland, and Norway: note that languages spoken in these latter countries were not represented in Google ngram).

Variables

Individualism. Hofstede (2001) provided individualism data at the national level and he expanded data later (Hofstede, Hofstede, & Minkov, 2010). The raw data for individualism were taken from Hofstede et al. (2010).

The country level individualism scores cannot be used in its raw form for language-based analysis. Because Google ngram data were for languages, not for countries except for British and United States based publications, it was necessary to compute language-level scores for individualism. Kashima and Kashima (1998) used the average of the cultural scores of countries that share the same language, when the unit of analysis was language. However, such an approach disregards population differences. The current analysis would be laden by a similar problem if individualism data were not to be weighted. Google ngram data depends on

the number of publications and the number of books published in different countries differs substantially for each language, rendering simple average score inappropriate. Thus, it was necessary to weight the individualism scores. The number of publications is a function of the population of a country besides other factors.

To check for the accuracy of weighting individualism scores by the respective populations of countries rather than the number of publications for each country, the correlation between population and number of publications was computed for the countries whose publication data were available (UNESCO, 2014). The correlation coefficient between population and publication was .987. Because the population data were not missing for any of the countries given by Hofstede but the publication data were, the individualism scores weighted by the population were used in the language-based analysis.

The individualism scores at the country level were weighted by the population of countries in 1970. The source file for population data was Penn World Table (Feenstra, Robert, & Marcel, 2013). The national individualism scores of various countries were weighted by their respective populations to compute individualism scores among people speaking a particular language. For English outside of US and Britain, countries used were Australia, Ireland, New Zealand, South Africa and Canada (77.7 % of population is native English speaker in Canada; thus 77.7 % of Canadian population in 1970 was used). For French, countries used were France, Belgium, and Canada (22.3% of population is native French speaker in Canada; thus 22.3 % of Canadian population in 1970 was used). For German, countries used were Austria, Germany, and Switzerland. For Chinese, countries used were China, Hong Kong, Singapore, and Taiwan. For Spanish, countries used were Argentina, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Mexico, Panama, Peru, Uruguay, and Venezuela. The weighted individualism scores for each language can be seen in Table 1.

The percentage of total pronoun use in written texts. Languages differ in terms of requirement to mention pronouns explicitly. Recall that ngrams in Google are standardized by dividing the number of occurrences of word(s) by the total number of words[2]. Therefore, it is possible to compare ngram data across countries. In Google ngram, parts of speech are tagged (Lin et al., 2012). Thus, it is possible to delineate, for example, between "change" as a verb and "change" as a noun. Pronouns were likewise tagged. The percentage of pronoun use in 1970 was directly retrieved from Google Books ngram database. The general pronoun datum for Russian was not available in Google Books ngram. The percentage of pronoun use in different languages and their variants can be seen in Table 1.

The relative use of first person singular pronoun use in written texts. Using the Google Books ngram data in 1970, the percentage of first person singular, i.e., "I, me", and plural, i.e., "we, us", pronouns were retrieved for American English, British English, Chinese, French, German, Hebrew, Italian, Russian, and Spanish. For English material published outside of US and Britain, the percentage given by the general English corpus is calculated net of material published in US and Britain. The resulting sample was composed of 10 languages and variants. Using these percentages, the relative use of first person singular pronouns over first person plural pronouns was computed by subtracting the total of "we and us" from the total of "I and me". The relative frequency of first person singular pronouns over first person plural pronouns can be seen in Table 1.

Results

The simple correlation between individualism and the percentage of total pronoun use in written texts was not significant, $r = .31$, $p = .42$ ($n = 9$-total pronoun data were not

---

[2] The absolute number of words was not available on Google ngram viewer.

available for Russian), whereas the simple correlation between individualism and the relative frequency of first person singular pronouns was significant, $r = .62$, $p = .05$ ($n = 10$)[3].

To competitively test the effect of the percentage of total pronoun use and the relative percentage of first person singular pronouns on individualism, a regression analysis was made, $R^2 = .64$, $p = .05$. The greater relative use of first person pronouns was related to higher levels of individualism, $\beta = .74$, $p = .02$, while total percentage of pronouns was not reliably related to individualism, $\beta = .34$, $p = .21$ ($n = 9$)[4].

Discussion

Two lines of research looking into the effects of language on cultural levels of individualism make different predictions regarding the relation between pronouns and individualism. On the one hand, priming studies showed that the use of first person singular pronoun ("I") within a particular language was related to the independent or individualist self and that the use of the first person plural pronoun ("We") was related to the interdependent or collectivist self (e.g., Brewer & Gardner, 1996; Gardner, Gabriel, & Lee, 1999; Na & Choi, 2009, Twenge et al., 2013, van Baaren et al., 2003). On the other hand, Kashima and Kashima (1998; 2003) found a strong correlation between the overall requirement to use pronoun and individualism.

The present study extended the findings from priming studies to cultural products in the form of written text. Specifically, the relative use of first person singular pronouns in texts written in different languages, but not the overall use of pronouns, was related to levels of individualism. This evidence was against the pro-drop hypothesis that suggests that the overall

---

[3] When "Other English" was excluded from the analysis, the results remained intact: the former $r = .34$, $p = .41$, and the latter $r = .64$, $p = .08$. Likewise, when the analyses were repeated with the country as the unit of analysis (Argentina, Belgium, Britain, , Chile, China, Colombia, Costa Rica, Ecuador, El Salvador, France, French Canada, Guatemala, Hong Kong, Israel, Italy, Mexico, Panama, Peru, Russia, Singapore, Spain, Taiwan, Uruguay, US, and Venezuela), the results parallel those found with language as the unit of analysis, : the former $r = .29$, $p = .17$, and the latter $r = .40$, $p = .05$.
[4] The results hold without the Other English group: the former $\beta = .76$, $p = .04$, and the latter $\beta = .36$, $p = .24$.

use of personal pronouns, but not the relative use of personal pronouns, is related to individualism.

What would, then, account for the strong correlation that Kashima and Kashima found (1998; 2003) between overall pronoun use and individualism? One may suggest that because the sample size in the present study was small, lack of significance should not be taken as evidence against pronoun drop hypothesis. Although the number of languages represented may be small, the amount of written material that was relied on includes 4-6% of all books ever published (Michel et al., 2011). Furthermore, if restricted range were the problem, one would expect the simple correlation between pronoun drop and individualism to be small. However, if the data given by Kashima and Kashima (1998) were to be used in the present analysis, the r would be .55 with language as the unit of analysis, and .81 with society as the unit of analysis. These correlations mirrors those found in Kashima and Kashima (1998): .64 for language based analysis, and .75 for society based analysis. Although it is not likely given such evidence, one cannot exclude the possibility that with a different sample of languages, the results may be different.

However, there are at least two alternative explanations that can account for the high correlation that Kashima and Kashima found between pronoun drop and individualism. A closer look at prodrop data in Kashima and Kashima (1998) reveals that the overwhelming majority of the languages that require the explicit use of personal pronouns are from the Indo-European, Germanic language family. Germanic languages have a lot in common besides the obligatory pronoun mention (Haspelmath, Dryer, Gil, & Comrie, 2005). Therefore, first, it is possible that Kashima and Kashima (1998) may have tapped into some other grammatical property by isolating Germanic languages through the pro-drop variable. Second, the societies speaking Germanic languages, which do not allow for drop of pronouns (Haspelmath et al., 2005), are perhaps already more individualist, and thus pronoun drop is only a proxy for these

other processes. Either way, the present evidence challenges the pronoun drop hypothesis, but lends support for the individual level experiments of first person singular vs. plural pronouns at the language level.

Sapir (1912) and Whorf (1956) proposed that patterns of thought are constrained by linguistic structure (see also Bloom, 1981 for similar claims). In this strong form, it is probably wrong (e.g., Hunt & Agnoli, 1991). However, its weak form that language influences cognition received support also in the present study. The lack of support for the proposed relation between pronoun drop and individualism does not imply that structural properties of language do not influence language or that pronoun drop do not influence any cultural variable, but it seems that pronoun drop is not related to individualism, whereas the relative use of first person singular pronouns over the plural pronouns is related to individualism.

References

Ayyash-Abdo, H. (2001). Individualism and collectivism: the case of Lebanon. *Social Behavior and Personality, 29,* 503-518.

Brewer, M. B., & Gardner, W. (1996). Who is this "We"? Levels of collective identity and self-representations. *Journal of Personality and Social Psychology, 71,* 83-93.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin, 125*, 47-63.

Gardner, W. L., Gabriel, S. & Lee, A. Y. (1999). I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science 10*, 321-326.

Feenstra, R. C., Robert, I., & Marcel P. T. (2013). The next generation of the Penn World Table. Retrieved from Http://www.ggdc.net/pwt

Haspelmath, M., Dryer, M.S., Gil, D. & Comrie, B. (2005). *The World Atlas of Language Structures.* NY: Oxford University Press.

Hofstede, G. (2001). *Culture's consequences* (2nd ed.). Thousand Oaks, CA: Sage.

Hofstede, G., Hofstede, G.J., & Minkov, M. (2010). Dimension data matrix. Retrieved from http://www.geerthofstede.nl/research--vsm/dimension-data-matrix.aspx

International Publishers Association (2013). *Annual report.* Retrieved from http://www.internationalpublishers.org/images/stories/MembersOnly/AnnualREPORTS/Annual_Report_201213.pdf

Kashima, E. S., & Kashima, Y. (1998). Culture and language: The case of cultural dimensions and personal pronoun use. *Journal of Cross-Cultural Psychology, 29,* 461-486.

Kashima, Y., & Kashima, E. S. (2003). Individualism, GNP, climate, and pronoun drop: is individualism determined by affluence and climate, or does language use play a role? *Journal of Cross-Cultural Psychology, 34,* 125-134.

Kemmelmeier, M., & Cheng, B. Y. (2004). Language and self-construal priming: A

    replication and extension in a Hong Kong Sample. *Journal of Cross-Cultural*

    Psychology, 35, 705-712.

Lau, I. Y. M., Lee, S., & Chiu, C. (2004). Language, cognition, and reality: Constructing

    shared meanings through communication. In M. Schaller & C. S. Crandall (Eds.), *The*

    Psychological Foundations of Culture (pp.77-100). Mahwah, NJ: Erlbaum.

Lin, Y., Michel, J., Aiden,E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic

    annotations for the Google Books ngram corpus. *Proceedings of the 50th Annual*

    Meeting of the Association for Computational Linguistics, 169–174.

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books Team, …Aiden,

    E. L. (2011a). Quantitative analysis of culture using millions of digitized books.

    Science, 331, 176–182.

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books Team, …Aiden,

    E. L. (2011b). Quantitative analysis of culture using millions of digitized books

    [Supplemental material]. *Science, 331*. Retrieved from

    http://www.sciencemag.org/content/early/2010/12/15/science.1199644/suppl/DC1

Morris, M. & Peng, K. (1994). Culture and cause: American and Chinese attributions for

    social and physical events. *Journal of Personality and Social Psychology, 67*, 949-

    971.

Na, J., & Choi, I. (2009). Culture and first-person pronouns. *Personality and Social*

    Psychology Bulletin, 35, 1492-1499.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and

    collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological*

    Bulletin, 128, 3-72.

Ross, M., Xun, W. Q. E., & Wilson, N. E. (2002). Language and the bicultural self.

Personality and Social Psychology Bulletin, 28, 1040-1050.

Sapir, E. (1912). Language and environment. *American Anthropologist, 14*, 254-279.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J.

Gumperz & S. C. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp.70-96).

Cambridge: Cambridge University.

Trafimow, D., Silverman, E. S., Fan, R. M., & Law, J. S. F. (1997). The effects of language

and priming on the relative accessibility of the private self and the collective self.

Journal of Cross-Cultural Psychology, 28, 107-123.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2013). Changes in pronoun use in American

books and the rise of individualism, 1960-2008. *Journal of Cross-Cultural*

Psychology, 44, 406-415.

UNESCO (1982). *An international survey of book production during the last decades*. Paris:

UNESCO.

UNESCO (2014). Number of titles by UDC classes. Retrieved from

http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=136&IF_Lan

guage=eng&BR_Topic=0

van Baaren, R. B., Maddux, W. W., Chartrand, T. L., de Bouter, C., & van Knippenberg, A.

(2003). It takes two to mimic: Behavioral consequences of self-construals. *Journal of*

Personality and Social Psychology, 84, 1093-1102.

Watkins, D., & Gerong, A. (1999). Language of response and the spontaneous self-concept: A

test of the cultural accommodation hypothesis. *Journal of Cross-Cultural Psychology,*

30, 115-121.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee*

Whorf. NY: Wiley.

Table 1.

The Values of the Variables Used in the Language Analysis

	Individualism	Total % of pronouns	The relative use of first person singular over plural pronouns
American English	91*	4.17	.173
British English	89*	4.13	.206
Chinese	19.97	2.67	-.233
French	71.62	5.26	-.033
German	66.09	7.33	-.032
Hebrew	54*	2.20	.116
Italian	76*	4.18	-.002
Other English	76.13	3.70	.091
Russian	39*	-	.102
Spanish	29.77	4.29	.021

*The individualism scores are unweighted country scores taken directly from Hofstede