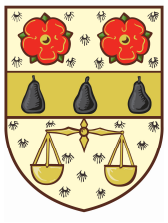
Nuffield's Working Papers Series in Politics



Examining survey translation validity using corpus linguistics

Jonathan Mellon, Nuffield College, Oxford

Email: jonathan.mellon@nuffield.ox.ac.uk

(Published 5th December 2011)

Abstract

Comparative public opinion research requires translation of survey instruments. A variety of translation approaches have been used, leaving most researchers unable to independently assess the quality of the translation. This paper looks at the different associations that keywords from surveys have across different languages by using 100-million word corpora collected from the internet in each language. Using these corpora, dissimilarity measures are constructed between languages allowing clusters of meaning to be identified quantitatively and examined qualitatively. Researchers can quickly apply this technique to examine whether the measures they use include unintended associations in some languages prior to empirical analysis. As well as providing a low-cost addition to current survey translation procedures, this paper uses a question from the World Values Survey to illustrate the method. I show that translation problems in several languages create confounding variables with this item.

Keywords: corpus linguistics, survey translation, World Values Survey, equivalence, post-materialism, measurement error

1 Introduction

Survey translation has been recognized as a potential cause of bias since the first multinational surveys [Verba, 1971]. However it has been difficult for researchers to assess whether translation problems will have an impact on their results unless they fluently speak all the languages their survey is conducted in. Without a way to directly assess the quality of translation, researchers do not know whether it is a major source of bias or a minor inconvenience.

One form of translation bias occurs when an important word in a question has different meanings across languages. Even if these differences are relatively subtle, the concepts that are tapped into in different languages could lead to confounding variables being introduced in some versions of the survey but not others. This will bias both population averages and estimates of individual level relationships involving the variable of interest. Even in those surveys that have been translated using best practice, the closest word in the target language to the original word in the source language may still not be equivalent as a word may have connotations and associations in one language that it lacks in another.

Corpus linguistics takes very large collections of text for a language (around 100 million words for each of nine languages in this paper) then returns the words most closely associated with another. By comparing these associations across languages it is possible to see whether the semantic associations of a word differ systematically when it is translated.

This paper outlines a method to assess how much the associations of a translated keyword differ across languages and what the content of those differences are. This paper compares the associations of the word *ideas* and its translations as taken from a post-materialism question in the World Values Survey [wvs, 2011]. I first compare the associations of *ideas* qualitatively for Chinese and English. The method is then expanded to quantitatively compare the meaning of *ideas* in 10 languages using cluster analysis. I find that the word *ideas* has systematically different associations in Western European languages compared to Chinese, Arabic, Polish, and Russian. The different meanings in these other languages are shown to predict the post-materialism item's relationships with other variables in the World Values Survey.

2 Meaning and equivalence

Cross-national research surveys rely on achieving equivalence of meaning for their questions across different languages and cultural contexts [Heath et al.,

2005]. Although there are many types of equivalence that have been defined [Johnson, 1998], it is useful for this paper to define it in terms of measurement error from response bias.

Response bias refers to any bias that causes the measured value, of an attribute of a survey respondent, to systematically deviate from the 'true value' of that attribute. This deviation can occur in several ways. In standard item response theory (IRT) a subject, i, in country k[1] gives a response, $\hat{\theta}_{ik}$, to an item, that measures their true score, θ_{ik}, with an error component ε_{ik} as shown in Equation 1.

$$\hat{\theta}_{ik} = \theta_{ik} + \varepsilon_{ik} \tag{1}$$

We assume that the 'true score' corresponds to the we are interested in such as happiness, social trust or post-materialism. Following Bollen's breakdown of model error [Bollen, 2002], we can decompose the measurement error into four further components as seen in Equation 2: a random component ε_{rik}, a confounding variable component, ε_{cik}, a uniform component ε_{uik} and a specification component ε_{sik}. These components are defined by their relationships with the true score, θ_{ik}, and the vector of all other attributes of an individual, i, in country k: X_{ik}. The vector includes both measured and unmeasured attributes.

$$\varepsilon_{ik} = \varepsilon_{rik} + \varepsilon_{cik} + \varepsilon_{uik} + \varepsilon_{sik} \tag{2}$$

Response bias generally includes the uniform and confounding components but these should be analysed separately as they are problematic in different situations and call for different responses.[2]

2.1 Error Component Definitions

The random error component is uncorrelated with either the true score: or any other individual attributes, measured or unmeasured, X_{ik}, as shown in Equation 3. However the magnitude of random error may vary according to X_{ik} or θ_{ik}. For instance lower educated respondents may be more likely to give random answers to a question.

$$Cov(\varepsilon_{rik}, X_{ik} + \theta_{ik}) = 0 \tag{3}$$

[1] This section refers to k as a country but it can equally refer to a language or cultural group.

[2] The random and specification components can also be affected by survey design and response errors but they are outside the scope of the paper.

The confounding variable component of the error is that part of the measurement error that correlates with other attributes of the respondent as expressed in Equation 4.

$$|Cov(\varepsilon_{cik}, X_{ik})| > 0 \tag{4}$$

Response biases that affect the confounding variable component can be problematic even within a single country as they mean that the variable is measuring something other than it is meant to.

The uniform error component, ε_{uik},is a constant systematic bias that is independent of any individual characteristics or their true score as shown in Equation 5.

$$\varepsilon_{uik} = a_k \tag{5}$$

The specification error component, ε_{sik}, is correlated with the true score: $|Cov(\varepsilon_{sik}, \theta_{ik})| > 0$. This is to allow for biases such as digit preference where respondents disproportionately report ages ending in zero or five [Carrier, 1959] or tendencies for the happiest people to under-report their true happiness. The specification error component is uncorrelated with all other individual attributes as shown in Equation 6.

$$Cov(\varepsilon_{sik}, X_{ik}) = 0 \tag{6}$$

2.2 Item Equivalence Conditions

Item equivalence can be defined as the situation where for any individual with true score θ'_i and a vector of all other attributes, X'_i, the probability density function of their measured scores, $pdf(\hat{\theta}'_i)$, is the same regardless of country.[3] Combining this with the definitions of the error components in the previous section gives the following four conditions for equivalence in Equations 7-10.

Firstly, the variance of the random error component should be constant, b, across all countries after accounting for the effect of covariates on the magnitude of ε_{rik} as shown in Equation 7.

[3]This is a relatively weak form of equivalence that merely requires that the item does not affect measurement differences across countries. A stronger form of equivalence would require that given θ'_i, $pdf(\hat{\theta}'_i)$ is the same regardless of k or X_i. This would require the same item equivalence conditions as the weak form but also that ε_{rik}'s magnitude is independent of X_{ik} and that ε_{cik} is zero. The effects of ε_{rik} and ε_{cik} are possible to account for within statistical models if they are constant across countries. Hence the weaker form of equivalence is also important.

$$\forall k : var(\varepsilon_{rik}) | X_{ik} = a \tag{7}$$

Secondly the covariance between the confounding error component and all properties of an individual respondent must be constant, a across countries k as shown in Equation 8.

$$\forall k : Cov(\varepsilon_{cik}, X_{ik}) = b \tag{8}$$

Bjornskov provides a good example of where this could fail to hold from the happiness literature: "the Russian and French translations of 'happy' both mean happy and lucky, implying that the word can have multiple meanings that make the scores less comparable across countries" [Bjornskov, 2010]. Any measured differences in happiness between English and Russian speaking countries could be driven by how lucky the Russian speaking respondents perceive themselves to be.

Thirdly the uniform component should be constant across all countries and individuals as shown in Equation 9.

$$\forall k : \varepsilon_{uik} = c \tag{9}$$

Finally, covariance of the specification component with the true score should be constant across all countries and individuals following Equation 10.

$$\forall k : Cov(\varepsilon_{sik}, \theta_{ik}) = d \tag{10}$$

Response bias affecting ε_{uik} and ε_{sik}'s equivalence across countries will affect population means derived from the instrument but won't strongly impact the instrument's correlations with other variables provided the true score's variance is still measured. These kind of biases are not problematic for variables measured on an arbitrary scale such as happiness or health as the uniform bias affects everyone's scores equally on measured scores. However the biases become problematic if they varies across surveys or countries as the scales become non-equivalent. This difference in uniform and specification bias is often referred to as differential item functioning (DIF). This kind of non-equivalence is common in translation. Bjornskov points out that translation of happiness questions can introduce this kind of bias: "the Danish translation 'lykkelig' is a stronger concept than the parallel notion of 'happy' in English and would arguably require more to achieve" [Bjornskov, 2010].

2.3 Detecting Confounding Bias

This paper assesses non-equivalence resulting from differences in confounding bias by looking at differences in the confounding error component. Although we can't observe the component's relationship with individual attributes directly we can observe their relationship to the measured values, $\hat{\theta}_{ik}$ which depends on the covariance of X_{ik} with both the true score, θ_{ik} and the error ε_{ik}. It is not possible to test all of X_{ik} for its relationship with the error. Firstly some attributes of a respondent will not be measurable. Secondly, in any real dataset there would be a high risk of Type I errors if all measured variables were tested as confounding variables of measurement. To actually test for confounding bias we should have good a-priori reasons, such as mistranslation, for suspecting a particular variable as a confounder, Z_{ik}. Provided we are testing a small number of potential confounders and we don't have strong theoretical reasons to suspect a systematically different relationship between the confounder and true value in different countries, we are justified in making the assumption in Equation 15 that the true relationship between θ_{ik} and Z_{ik} is constant across countries. From this it follows that any significant observed difference in the relationship between $\hat{\theta}_{ik}$ and Z_{ik} is due to violation of the confounding error component's equivalence.[4]

3 Previous Approaches to Assessing Equivalence

There has been a lot of attention paid to identifying and correcting DIF where there is a difference in the uniform or systematic response bias across groups. DIF can sometimes be corrected using anchoring vignettes [King et al., 2004]. For instance if the word "healthy" had a stronger meaning in English than in Chinese we could correct for this by measuring the respondents' ratings of a common situation such as:

> Bob has difficulty walking more than 100 metres and sometimes misses work due to this. Using this scale from 1 - 10 where 1 is not healthy at all and 10 is completely healthy how healthy would you say Bob is?

If the word healthy is much stronger in Chinese than English then we would find that Chinese respondents rate Bob's health as better than English respondents do. After measuring this difference it is possible to rescale the responses about one's own health to an equivalent scale between countries

[4] For a formal derivation of this see Appendix B.

[King et al., 2004]. Importantly this rescaling is neutral regarding the cause differences in uniform and systematic response bias. It will work whether the bias is caused by differences in expectations of good health or a difference in the strength of the meaning of the words.

For constructs that are measured using multiple indicators several statistical approaches have been proposed for achieving equivalence including multigroup confirmatory factor analysis [Nye et al., 2008], multiple indicators multiple causes [Pérez, 2011], and IRT latent variable models [Stegmueller, 2011]

These approaches all aim at addressing DIF. However none of the approaches can directly detect or correct for confounding response bias. Instead they either make zero confounding response bias an assumption [King et al., 2004, Stegmueller, 2011] or point to high inter-item correlations as evidence against its presence [Pérez, 2011]. This is understandable as detection of confounding response bias requires some a-priori reason to suspect a particular confounding variable in the measurement. Testing all available variables would risk finding many false positives due to multiple testing or many false negatives if an adjustment was made for multiple testing. Insofar as the assumption of no confounding response bias is false, the measurement models previously suggested are biased due to misspecification and are incorrectly estimating DIF as some of the difference will be explained by omitted variables.

Harkness, Villar, and Edwards describe team translations as current best practice for achieving equivalence [Harkness et al., 2010]. Willis et al. outline one such approach: the Translation, Review, Adjudication, Pretesting and Documentation model (TRAPD) [Willis et al., 2010]. Several translators produce drafts which a larger team then reviews. Once a draft has been agreed, pretesting through in-depth interviews with members of the target population identifies comprehensibility problems and errors in translation.

A commonly used assessment tool is back translation where the translated text is translated (by a different translator) into the original source language so the original document can be compared to it. Harkness and Schoua-Glusberg describe back translation as a "primitive" tool that will catch some errors but miss many others as it only indirectly accesses problems in the target language [Harkness, 2003]. In particular, back translation cannot identify situations where a translation is as close as possible to the source but still non-equivalent. For instance with the Russian happy example there may not be a better Russian translation of happy, so back-translation will not find a problem with the translation even though it is bringing in the construct of luck as well.

Current translation techniques such as TRAPD aim at improving the translation of a survey prior to conducting it. However this is not useful for researchers who use existing multinational survey data; particularly as existing surveys were translated using a wide range of methods. Currently, researchers must simply hope that a survey adequately conveys the intended meaning across different languages, or else avoid using any survey not translated using best practices (which are far from universally agreed upon). The latter approach is overly wasteful given that, even without implementing the gold standard, many questions probably convey meaning across different languages. Avoiding surveys translated using older approaches would exclude most of the social science data currently available, which is unrealistically demanding for scholars. However a researcher has no way of knowing which questions they can safely use without making false inferences. Most research implicitly follows the former option and uses whichever data are available to address their substantive interests regardless of the methods used to translate the survey. The overview of current practice shows that there is a need for tools that can assess translation problems in existing surveys.

4 Assessing Meaning using Corpus Linguistics

Corpus linguistics looks at the structure and meaning of language by examining large bodies of naturally generated text from a language. This paper uses two techniques developed within corpus linguistics: concordance and collocation. Concordances are examples of sections of the corpus that use a given word (chosen by the researcher) showing the context in which the word is found. These are useful for seeing examples of how a word is actually used in a language. Concordance lines are useful for seeing the exact context a word is used in but are impractical for an overview of a word's usage as there are often several thousand examples of a word. It is difficult to generalise about a word's usage and more subtle patterns may not be recognised.

The technique of collocation is more useful for finding these patterns. Two words are collocations of each other if they are found together more often than chance. A number of statistical techniques can assess the strength and significance of the association between the two words [Dunning, 1993, Thanopoulos et al., 2002]. The context a word is found in gives us insight into the way the word is used and its meaning.

Using the collocations of a word to assess its meaning has a strong justification in terms of how vocabulary is acquired. Although the precise method by which children acquire the meaning of words is disputed [Bloom, 2002], it is generally agreed that children infer the meaning from the context the

word is used in [Bloom, 2002]. A large part of this context are the other words found near the word being learnt [Bloom, 2002, Chaffin et al., 2001]. So if a word in one language is found surrounded by a very different set of collocations to those surrounding a word in a second language, it is implausible that they can mean the same thing as the mechanism would lead people acquiring the vocabulary to infer a different meaning.

For the purposes of this paper, being 'found together' is defined as the collocation being found within 10 words to the left or right of the keyword. This span is wider than that used in many corpus studies. There are two reasons for this choice. First, the collocations need to be comparable across languages and a shorter span would mean the collocations found for a language would be more dependent on its syntax. For instance German generally puts the verb at the end of a sentence so a short span would bias towards finding certain parts-of-speech in the collocations. A longer span removes the effect of syntax by including all of the sentence. The second reason is that when subjects are faced with novel words they reread and scan a wide section of text to try and determine the word's meaning [Chaffin et al., 2001]. This suggests that words found within the wider span are likely to be relevant to the meaning of a word.

Corpus software allows a researcher to enter a word of interest and quickly retrieve a list of the word's most associated collocations. Patterns in a word's collocations can be extremely useful in understanding its meaning. In this paper the Leeds Collection of Internet Corpora [Sharoff, 2011] is used to retrieve collocations in each language, listed in order of the strength of association between the collocation and keyword as calculated by a log-likelihood ratio statistic. The corpora cover 12 languages: Arabic, Chinese, English, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian, and Spanish with between 100 million and 200 million words in each derived from between 35,000 and 40,000 documents [Sharoff, 2011]. These properties make the collection of corpora ideal for examining translation in multinational surveys. Custom corpora would make for a more customisable approach but would make translation evaluation impractical for most researchers as it would be necessary to collect many large corpora before any analysis would be possible.[5]

[5]Ideally there pre-constructed corpora would be available to download but it is not currently legal to distribute corpora due to copyright restrictions on the works contained within them. Corpora only using open source texts are available but these reflect a highly skewed sample of a language. Out of copyright works can also be used but these are always many decades old and won't reflect modern language usage.

5 Comparing Chinese and English Associations for a Post-Materialism Item

In the most simple case corpus linguistics can be used to examine the differences in the meaning of a survey question across a pair of languages by comparing their collocations. As an example I take a question from the World Values Survey on post-materialism. Post-materialist values "emphasize self-expression and the quality of life" [Inglehart and Welzel, 2005]. Inglehart argues that individuals are more likely to adopt these types of values if they do not have to worry about survival so post-materialism should increase as societies develop and become richer [Inglehart and Welzel, 2005]. The question reads:

> In your opinion, which one of these is most important?
>
> And what would be the next most important?:
>
> - A stable economy
> - Progress toward a less impersonal and more humane society
> - Progress toward a society in which Ideas[sic] count more than money
> - The fight against crime

The first and last options are considered materialist responses and the second and third are post-materialist. In this question there are a number of words that are important for a respondent's interpretation of the question. In this section I look at the word *ideas* from the third option (I consider the other words in appendix D). The word *ideas* is an abstract word but needs to be given a concrete interpretation by a respondent in order for them to assess its value relative to money. As a result it should be a particularly difficult word to find a translation with equivalent meaning for in another language. The collocations for each language are retrieved from the Leeds Collection of Internet Corpora website [Sharoff, 2011].

5.1 Translate and clean

In order to compare the collocations the keyword translations have in different languages it is necessary to translate them into a common language.[6]

[6]Bing Translator includes an Application Programming Interface that allows it to be automatically called by R scripts which can greatly speed up the process compared to

This paper uses Microsoft's Bing Translator online service [Microsoft, 2011] to translate each list of collocations into English allowing a direct comparison of them.[7] Because we are relying on lists of up to 100 words for each language these errors should not strongly affect the patterns of similarity between languages even if some errors occur. Using a full list of collocations spreads the risk of mistranslation compared to relying on a single translation of the keyword. The lists are cleaned of semantically non-important function words such as pronouns and conjunctions as these are not important to the comparison of meaning and are more related to the grammar of a language which would bias the comparison.

5.2 Differences in the Meaning of *ideas* between Chinese and English

The collocations of *ideas* for Chinese and English are shown in Table 5.2. The English collocations fit reasonably well with how *ideas* are usually understood. They list mainly focuses on expression (e.g. *express*, *discussion*, and *share*) and generation (e.g. *creative*, *develop*, and *design*). The Chinese list of collocations has a very different set of themes. The two main themes are teamwork/team spirit (*carry forward*, *dedication*, and *playing*) and community (*culture*, *community*, and *tradition*). From this comparison it seems very unlikely that *ideas* and the Chinese translation of it, used in the World Values Survey, have equivalent meanings. These results suggest that there may be a confounding bias in the Chinese survey that could undermine equivalence.

entering the lists manually. This step is likely to introduce some errors into the process as Bing Translate will sometimes make incorrect translations. However the errors it introduces are likely to make the two lists look less similar (i.e. two equivalent words are not translated equivalently) rather than more so. By contrast we can be reasonably confident that collocations that appear on two lists are correctly translated, as the probability of mistranslating a word into one that happens to be on a specific list is extremely low.

[7] The original intention was to use the Google Translate API for this method. However Google announced that they were deprecating the API by the end of 2011 and would begin charging for a new version after this [Feldman, 2011]. The analyses in this paper were replicated using Google Translate and the results do not differ substantially.

English		Chinese	
Collocation	Joint Frequency	Collocation	Joint Frequency
idea	1577	people	5016
people	1079	substance	1409
share	672	culture	1796
student	750	people	1316
book	760	life	1789
exchange	459	status	1151
help	688	body	785
develop	574	team	753
concept	432	reflect	856
information	651	dedication	603
express	358	innovation	806
time	683	thinking	1007
creative	303	china	1599
suggestion	284	humanities	681
discuss	347	science	869
learn	427	community	1101
mind	356	carry forward	488
provide	427	body	878
activity	369	pursue	713
experience	377	dedicate	526
generate	260	student	1120
design	377	culture	690
project	377	implement	486
look	438	age	725
lot	347	free	716
try	382	power	649
include	449	focus	583
teacher	308	pressure	596
discussion	279	promote	366
explore	240	performance	667
question	369	collapse	433
write	349	education	795
business	364	pillar	366
create	337	time	1194
communicate	206	ability	691
free	326	playing	372
change	383	inspire	343
		mankind	589
		people	649
		patient	500
		food	271
		wealth	450
		awareness	490
		people	680
		developing	764
		revolution	512
		torment	420
		play strong	251
		tradition	532

Table 1: Collocations for *ideas* translations in English and Chinese

6 Comparing Meaning of *ideas* Translations across 10 Languages

This method of comparing translations using corpus linguistics can be generalized beyond two languages. Comparing two lists of collocations can be instructive for understanding the differences between two languages. However when looking at several languages it becomes difficult to understand the overall structure of similarities between all the languages. For 10 languages such a comparison would involve simultaneously considering 45 two-way comparisons. In order to examine the structure we need to analyse the different meanings shown in the collocations quantitatively. This section expands the method to compare the meaning of *ideas* across 10 languages using cluster analysis to highlight possible cases of non-equivalent meaning. This article compares responses in 10 languages[8] : English, German, French, Spanish, Portuguese, Russian, Polish, Chinese, Arabic, and Italian.[9] The cluster structure is then given meaning by looking at which collocations are shared between languages in a cluster.

6.1 Construct Dissimilarities Matrix

The collocation lists for each language are downloaded from the Leeds Collection of Internet Corpora as before and translated into English using Bing translate. The translations of *ideas* used in the World Values Survey and details of each corpus are shown in Table 6.1. Once the lists of collocations are in a common language they can be compared statistically. Several collocations may be translated into the same word in English in which case their frequencies are combined to represent a single collocation. For every collocation a joint frequency is given that shows how many times it turns up in the corpus with the keyword. In the analysis of *ideas* in Table 5.2, the word *ideas* has a joint frequency with the collocation *express* of 358 meaning they appear together 358 times in the Corpus.

For each list, these joint frequencies are converted into percentages of the total joint frequencies in the list so that the sum of the joint frequencies in a list sums to 1. In this example there were 17,364 valid collocations with

[8]This paper uses the ISO 639-1 language codes to identify languages in figures and tables. The relevant ones in this paper are English (en), German (de), French (fr), Spanish (es), Portuguese (pt), Russian (ru), Polish (pl), Chinese (zh), Arabic (ar), and Italian (it).

[9]Greek also has an available Corpus but it was not included as it does not include part-of-speech tagging. Japanese was also available but not included due to difficulties integrating part-of-speech tagging queries within R.

Language	Corpus Size	Ideas Translation	Ideas Frequency	Ideas Percentage of corpus
English	181,376,006	Ideas	58,846	0.0324%
Chinese	281,660,631	精神	52,270	0.0186%
Arabic	165,674,718	الأفكار	7,938	0.0048%
French	263,727,981	Idées	82,268	0.0312%
German	255,595,925	Ideen	41,311	0.0162%
Italian	155,569,389	Idee	42,635	0.0274%
Polish	86,605,024	Ideały	2,530	0.0029%
Portuguese	129,920,567	Idéias	20,625	0.0159%
Russian	198,509,029	идеалы	5,754	0.0029%
Spanish	145,572,631	Ideas	52,766	0.0362%

ideas. This means that *express* makes up 2.1% of total collocations in the list.

The next step is to calculate a measure of dissimilarity between every pair of languages. The lists of collocations are first combined so that a joint frequency percentage is given for every collocation that appeared on either list for both languages. This means that any collocation that only appears on the list of the first language is given a joint frequency of zero in the second language. The differences between the lists are then calculated by taking the absolute difference between the joint frequency percentages for each collocation and summing the differences. The total is then divided by two to give a score between 0 and 1.

The difference score for two languages is summarized in equation 11 where f_{xi} is the joint frequency percentage for the ith collocation on the combined list of collocations for language x, and n is the number of collocations on the combined list.

$$\frac{\sum_{i=1}^{n} |f_{1i} - f_{2i}|}{2} \tag{11}$$

This equation is applied to each pair of languages and the resulting scores are entered into a dissimilarity matrix. The dissimilarity score between two languages can range from 0 (if their collocation lists were identical) to 1 (if the lists had no collocations in common).

A list of collocations for *ideas* was retrieved in each language and a matrix of dissimilarities was constructed based on a comparison of each list. The dissimilarities matrix for *ideas* is shown in Table 2.[10]

[10]To look at whether the translation engine affected the dissimilarities matrix, the *ideas* matrix generated by Bing Translate was correlated against that translated using Google

	en	zh	ar	fr	de	it	pl	pt	ru	es
en	0	0.87	0.83	0.74	0.66	0.70	0.92	0.68	0.86	0.79
zh	0.87	0	0.96	0.87	0.91	0.91	0.95	0.87	0.92	0.92
ar	0.83	0.96	0	0.90	0.82	0.82	0.92	0.84	0.86	0.89
fr	0.74	0.87	0.90	0	0.78	0.68	0.89	0.69	0.90	0.70
de	0.66	0.91	0.82	0.78	0	0.66	0.90	0.78	0.87	0.77
it	0.70	0.91	0.82	0.68	0.66	0	0.91	0.65	0.78	0.56
pl	0.92	0.95	0.92	0.89	0.90	0.91	0	0.93	0.81	0.92
pt	0.68	0.87	0.84	0.69	0.78	0.65	0.93	0	0.80	0.63
ru	0.86	0.92	0.86	0.90	0.87	0.78	0.81	0.80	0	0.86
es	0.79	0.92	0.89	0.70	0.77	0.56	0.92	0.63	0.86	0

Table 2: Dissimilarities matrix for *ideas* translations

6.2 Cluster Analysis

This paper analyzes the dissimilarity data using cluster analysis. Cluster analysis refers to a wide variety of methods for grouping a number of objects into mutually exclusive categories based on the similarity between the objects in a category. [11]

A relatively simple hierarchical algorithm called DIANA is used in this paper [Kaufman and Rousseeuw, 1990]. All objects are initially considered as a single cluster. The object with the highest average dissimilarity to all other objects is removed from the cluster and used as the basis for a new cluster. Each object in the initial group is then examined to see whether it is closer, on average, to the new cluster or the original. If the object is closer to the new group it is moved into the new cluster. This process is repeated until no object in the initial group is closer to the new group than its own. This process of splitting is applied iteratively to each of the new clusters until all clusters consist of a single object.

The cluster analysis of the translations is intended to provide a guide to where a researcher should look for potentially problematic translation issues but shouldn't be seen as evidence of translation problems by itself.[12]

Translate. They correlated at r = 0.96 after excluding zeroes which would have further increased the correlation showing that the translation method used is not greatly skewing the findings.

[11] Many clustering algorithms requires the user to specify the number of clusters in advance. These approaches are not particularly useful for translation as we do not necessarily know how many (if any) languages will have translation problems in advance of the analysis. For this analysis hierarchical approaches are more useful.

[12] Several different forms of hierarchical clustering were performed for the analyses in this paper. Different techniques showed no differences that would affect the substantive

The importance of the clustering structure can be tested by looking at the most common collocations shared within a cluster. If the collocations shared in one cluster are qualitatively different to those in another cluster then that division is probably meaningful in dividing the meaning of the keywords across languages.

The overall picture from using these techniques on the *ideas* translations is as follows. The Western European translations bring in concepts of innovation and sharing. Polish and Russian translations bring in primarily political and moral concepts. The Chinese translation is associated with community and dedication to a goal as well as some association with mental wellness. The Arabic translation seems to be mainly negative with some religious (or anti-religious) connotations.

The result of applying divisive clustering analysis to *ideas* is shown in Figure 1. The clustering structure shows Arabic, Chinese, Polish, and Russian are all split into their own clusters early on although Polish and Russian are more similar. Within the Western European languages the Germanic languages of English and German are split off first with the Romance languages more closely clustered.

Firstly we can examine the clustering within the European languages. Table 6.2 shows the collocations shared within English and German compared to those shared within French, Spanish, Portuguese, and Italian. Although the lists contain quite a few different words the concepts they are related to are fairly similar. Both lists include many words related to creation and development (e.g. creative, develop, express) as well as some related to transmission (e.g. give, exchange). The split seems substantially unimportant and the two lists do not support the hypothesis that the word *ideas* means something very different between the two sets of languages.

A considerably different picture emerges when comparing the Polish and Russian cluster to the Western European cluster as a whole as shown in Table 6.2. Whilst the Western European languages follow the pattern previously seen, the Polish and Russian collocations more closely relate to ideology and morality (e.g. faith, moral, justice, freedom, believe).

Table 5 shows the collocations of *ideas* for Arabic. Three themes come out of the Arabic list. There are several negative words such as “suspicious”, “sneaky”, and “misconceptions”. There are also several religious terms such as “Al-Isra” and “Miraj” (two parts of the night journey taken by the prophet Mohammed) and political terms such as “liberalism” and “leader”. These themes combined with an examination of Arabic concordance lines lead to a

conclusions reached so only the results of divisive clustering using the average method are shown.

French, Spanish, Portuguese, and Italian		English and German	
collocations	number of languages shared in	Collocations	number of languages shared in
All	4	about	2
Also	4	all	2
By	4	also	2
Do	4	at	2
Duty	4	by	2
express	4	can	2
Give	4	come	2
Have	4	concept	2
Idea	4	creative	2
If	4	develop	2
More	4	find	2
New	4	from	2
Not	4	give	2
power	4	have	2
Say	4	here	2
without	4	idea	2
world	4	if	2
between	3	make	2
develop	3	more	2
exchange	3	new	2
Go	3	not	2
Good	3	on	2
knowledge	3	one	2
Lot	3	other	2
Man	3	own	2
Own	3	they	2
person	3	time	2
share	3	what	2
thing	3		
Want	3		
When	3		

Table 3: Collocations shared within clusters of languages for *ideas* within western europe

Polish and Russian		English, German, Spanish, Italian, Portuguese, French	
collocations	number of languages shared in	collocations	number of languages shared in
another	2	all	6
at	2	also	6
Christian	2	by	6
freedom	2	give	6
from	2	have	6
great	2	idea	6
ideal	2	if	6
justice	2	more	6
man	2	new	6
moral	2	not	6
not	2	develop	5
only	2	do	5
political	2	express	5
same	2	own	5
sublime	2	your	5
such	2	about	4
value	2	come	4
what	2	concept	4
world	2	duty	4
		exchange	4
		good	4
		lot	4
		man	4
		on	4
		other	4
		power	4
		say	4
		share	4
		time	4
		want	4
		what	4
		without	4
		world	4

Table 4: Collocations shared within Russian and Polish compared to those shared between Western European language for *ideas*

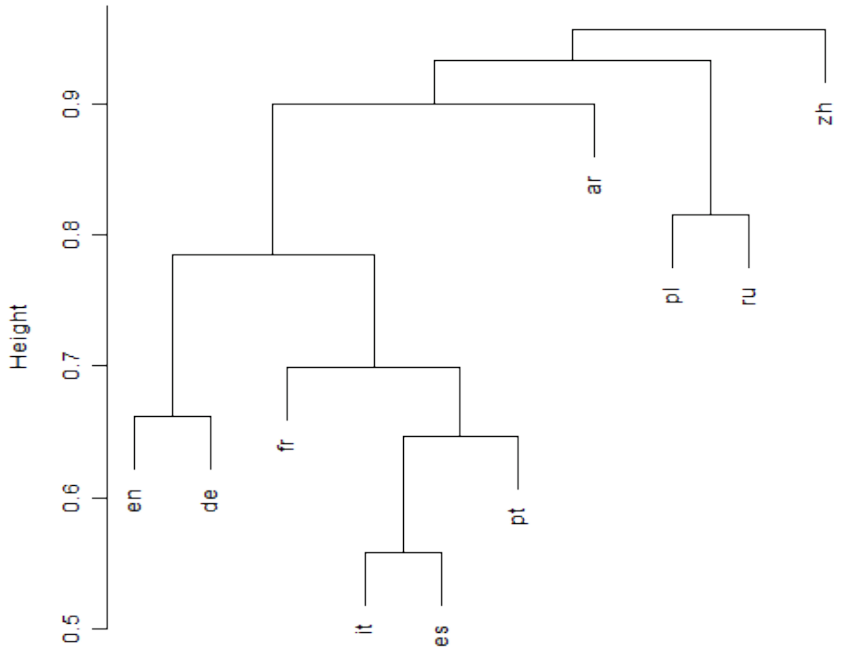


Figure 1: Hierarchical clustering dendogram of *ideas* dissimilarities

tentative conclusion that the Arabic translation of *ideas* is most commonly used to describe contentious ideological or religious concepts. For instance one concordance (translated using Google Translate) reads:

> Fascism, and Nazism, we must also continue to protect [ideas] from the [ideas] of the moment ... If we are demanding the protection of our political...

whilst another reads:

> even if the [ideas] are not consistent with the values and customs of [ideas], and destruction, not reconstruction. Everyone has to adopt for himself what he wants of the...

Although the automatic translation is poor, the context that the Arabic translation of *ideas* is used in is clearly political and negative.

6.3 Evaluation of Other Words in Post-Materialism Question

The words *ideas* is not the only important words for the interpretation of the postmaterialism question. The following words were also checked for

Collocation	Joint
about	1203
al-isra	90
any	194
because	133
between	456
can	229
during	180
each	554
even	210
from	4245
group	122
idea	242
ideas	687
if	504
including	144
leader	91
liberalism	110
life	120
lot	133
may	445
mean	197
miraj	90
misconceptions	77
model	90
more	146
new	210
no	1320
non	228
nor	300
not	154
occasion	104
outlooks	90
political	112
salutes	90
scandalous	90
sneaky	90
society	127
some	498
subtract	80
such	261
suspicious	90
these	1581
thinker	91
those	358
thought	100
views	154
what	817
when	191
where	204
which	2321
world	146

Table 5: Collocations of *ideas* in Arabic from the Leeds Collection of Internet Corpora

differences in meaning in the question: *economy*, *progress*, *impersonal*, *humane*, *fight*, and *crime*. None showed a clear difference in meaning (although there were some borderline cases where the word was not found frequently enough within the corpus to generate a stable list of collocations). For the full analysis of the other words see appendix D.

6.4 Summary of Linguistic Findings for Post-Materialism Question

In light of this discussion, the important findings of the linguistic analysis are:

- In the six Western European languages the word *ideas* reflected notions of innovation and sharing concepts;
- In Arabic the translation of *ideas* referred more to anti-religious feelings and subversive concepts;
- In Chinese the translation of *ideas* is related to community, dedication, and team spirit;
- The Polish and Russian the *ideas* translations both bring in political and religious concepts.

7 Survey Analysis of Associations between the Question and Other Concepts

The linguistic analysis of the translations in this paper suggest the word *ideas* could tap into different meanings across languages and lead respondents to answer using different attitudes. To demonstrate the potential problems that this could cause in empirical analysis, this section looks at some of the differing associations found between the *ideas* item and other variables in the World Values Survey and European Values Survey across the different clusters of languages found in Section 6 .

These analyses turn the *ideas* item into a binary variable which is coded as 1 if the respondent chooses the *ideas* option as either their first or second choice. Respondents are coded according to which of the four language clusters identified in the cluster analysis of *ideas* they fall into: Western European, Polish and Russian, Chinese or Arabic. People speaking other

languages or who are minority language speakers in their country were excluded from the analysis.[13]

A logit model is then fitted to look at the effect of the potential confounding variable and whether it has a significantly different effect in different clusters of languages by interacting it with a dummy variable for language cluster. The Western European group is used as the reference group in each model. A fixed effect is included for each survey so that country and time specific changes do not affect the results. Demographic controls were not included in the models as the hypothesis is that both the confounding variable and the *ideas* item tap into the same latent variable. If demographic variables are correlated with both it is most likely through the latent variable so including them in the regression could make the association less clear by reducing the variance in the underlying attitude.

The specification is as follows:

$$\begin{aligned} logit(p(ideas)) = \beta_0 + \beta_1 Confounder + \beta_2 Chinese * Confounder + \\ \beta_3 Arabic * Confounder + \beta_4 Polish/Russian * Confounder + \\ \beta_i survey_i + ... + \beta_n survey_n \end{aligned}$$

Because the questions asked in the World Values Survey vary over time the regression models have different numbers of observations depending on the independent variables used in the analysis. The table in Appendix C shows the number of observations available for each of the independent variables used here.

7.0.1 Nationalism

The Chinese translation of *ideas* is related to community and dedication and the Russian/Polish translations were related to political ideas. Both of these are likely to create an association with nationalist feelings compared to the Western European translations. To test this hypothesis two questions from the World Values Survey were used. The first asked "How proud are you to be [Your nationality]?" with the responses coded on a 4 point scale from "Very proud" to "Not at all proud" and the second asks "Of course, we all hope that there will not be another war, but if it were to come to that, would

[13]The earlier World Values Surveys and European Values Surveys did not record the survey language in the data. For these surveys, language was imputed if more than 90% of respondents in that country answered in a single language in years where it was recorded. For instance, Canadian surveys are not included for earlier years as they do not record whether they were conducted in French or English.

you be willing to fight for your country?" with the respondent answering yes or no.

The results for this are shown in Table 6. For both of the variables the interaction with Polish/Russian language and the interaction with Chinese language are positive showing that there is a more positive relationship between nationalist concepts and the *ideas* items in these surveys as predicted compared to Western European countries. In addition both questions also show a positive interaction with the Arabic language. This relationship was not predicted *a-priori* but may make sense in the Arab political context as secular nationalist movements are often contrasted with Islamic ones [Tibi, 1997].

Table 6: Logit models predicting probability of picking "ideas" option first or second

	(1) Importance of tradition	(2) Willingness to fight for country	(3) Importance of God	(4) Proud of nationality
Importance of tradition	-0.0575***			
	(0.0117)			
Willingness to fight for country		-0.190***		
		(0.0508)		
How important is god in your life			-0.0405***	
			(0.0108)	
How proud of nationality				-0.257***
				(0.0437)
Interaction of confounder with Chinese	0.139***	0.563***	-0.0372***	0.434***
	(0.0117)	(0.0508)	(0.0108)	(0.0437)
Interaction of confounder with Russian/Polish	-0.0417*	0.268**	0.0383**	0.365***
	(0.0169)	(0.0965)	(0.0147)	(0.0697)
Interaction of confounder with Arabic	0.0111	0.272***	0.00840	0.352***
	(0.0359)	(0.0745)	(0.0436)	(0.107)
Constant	-0.785***	-0.889***	-0.701***	-0.0401
	(0.0484)	(0.0345)	(0.0788)	(0.162)
Observations	28561	87418	105829	106383

Standard errors in parentheses. Fixed effects for each survey are included but not shown.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

7.0.2 Traditional Values

Another predicted association is that Chinese respondents who value tradition will be more likely to favour ideas. To test this we can look at an item asking how like the respondent a hypothetical person is. In this case the person is described as: "Tradition is important to this person; to follow the customs handed down by one's religion or family." By contrast none of the other language groups were predicted to share this association, from the linguistic analysis.

As expected the interaction between the importance of tradition and Chinese language is strongly positive as can be seen in Table 6. By contrast with the nationalism items, the relationship is actually significantly negative for the Russian/Polish language interaction although the relationship is not strong.

7.0.3 Religiosity

Both the Russian/Polish translations and the Arabic translations of *ideas* link to religious concepts. With the Russian/Polish referring to religious terms and the Arabic referring to anti-religious ones. The clear prediction from this was that Arabic would show a negative relationship with religious belief and Polish/Russian would show a positive relationship.

This relationship was difficult to test for all the language clusters due to a lack of variance in religious belief within languages. When looking at the question "How important is God in your life?" measured on a 10-point scale, over 90% of Arabic speaking respondents responded with a score of 10 (Very important). Over 50% of Polish speaking respondents also responded with the highest score of importance. By contrast, 52% of Chinese speaking respondents gave a score of 1 (Not at all important) to the question. These variances are reflected in the high standard errors of the coefficients in Table 6. The Russian/Polish cluster shows a significantly more positive relationship with *ideas* than the Western European cluster (reflected in the positive coefficient) as predicted but the confidence interval on the Arabic coefficient is too large to meaningfully interpret.[14]

7.1 Summary of Method

The method in this paper can be summarized in the following steps:

[14]If the Russian and Polish languages are analysed separately the confidence intervals on the Polish interaction also become very large so the coefficient is being primarily driven by variation in the importance of God among Russian respondents.

1. Identify a keyword within the survey question of interest in the first language. Identify the equivalent keyword in translations of the same survey.
2. A list of words most strongly associated with a keyword (its collocations) are retrieved in each language from the Leeds Collection of Internet Corpora . These show how the keyword is used in each language.
3. The collocations are then translated into a common language using translation software (Bing Translate in this paper) [Microsoft, 2011].
4. The lists of collocations are stripped of semantically unimportant words to avoid biases introduced by varying grammatical structures in different languages. For instance pronouns are rarely used in Spanish.
5. A dissimilarity score is calculated between each pair of languages based on how different their lists of collocations are. The scores are entered into a dissimilarity matrix.
6. The dissimilarity matrix is analyzed by mapping the positions of languages in relation to each other using multidimensional scaling. Languages in which the keyword has similar meaning are found using cluster analysis.
7. The clusters are interpreted semantically by looking at the collocations shared between languages within a cluster. This analysis is used to understand where potentially important differences in survey question meaning are likely to be present.
8. The meaning of the word can be cross-checked against translated examples of its use in concordance lines.
9. Finally the above analysis can be used to make predictions about different confounding associations the question of interest could be expected to have across different languages. These predictions can then be tested against other variables in the survey to assess the impact of any translation problems.

8 Discussion

The case study in this paper demonstrates the ability of corpus linguistic methods to detect translation problems in important words in survey questions. In this case the translation problems introduce substantial confounding

variables in some countries as shown by the analysis of the World Values Survey. Several different confounding variables are found that indicate that the equivalence condition for the confounding error component, ε_{cik}, in Equation 8 does not hold for the post-materialism item. This could undermine analysis done with this instrument leading to false inferences or masking the true relationships with other variables of interest. This tool will help researchers using cross-national data to avoid such errors and have greater confidence in the validity of the data they rely on.

The method in this paper opens up several lines of future research. Most importantly the translation of existing surveys should be examined and the implications for empirical results across social science should be tested. In particular, items that attempt to measure abstract concepts should be examined for translation problems. A good example of this is the trust literature which attempts to measure a fairly elusive concept with a variety of questions.

The method should become a standard diagnostic when performing analysis of cross-national survey data. The process in this paper will be released as an R package that automatically implements all the steps (with the exception of the analysis of survey data) described in section 7.1 for a set of translations entered by a user. This provides a simple low cost tool for assessing translation validity before making inferences.

The method should be extended by adding more corpora in more languages. The only requirement is that its speakers must have a sufficiently large internet presence to generate a corpus with enough variety of documents. Internet corpora have been created in languages from Kiswahili [De Pauw et al., 2009] to Farsi [Pendar and Sharoff, 2009] so it will certainly be possible to expand this method to many more languages than those covered by the Leeds Collection of Internet Corpora.

Another extension would be to collect larger corpora as low frequency words are currently difficult to examine using this method. The Google Ngrams project [Michel et al., 2011] could be one source of data in the future but is currently limited to seven languages and can only provide collocations one word to the left or right of a keyword which would bias the dissimilarity scores in favour of grammatically similar languages.

The potential to systematically identify the presence of confounding bias opens up new possibilities for measurement models. It may be possible to create models which can correct for both differential item functioning and differential confounding biases and get considerably closer to measurement equivalence in survey research.

Translating survey instruments is unavoidable if we want to address the most important questions in social science and make valid generalisations about the world that extend beyond English speaking countries. The case

study in this paper shows how translation problems on a single keyword can seriously threaten the validity of a survey instrument. The method put forward in this paper uses corpus linguistics to directly compare the meanings of a word translated across languages according to how it is used in natural situations. The method provides a tool for researchers to distinguish those survey questions with serious translation problems from those which manage to convey the same meaning across languages. Such an approach will allow scholars to make a more informed decision about which countries and questions to include in their analyses.

9 Bibliography

Jonathan Mellon is a DPhil candidate in Sociology at Nuffield College. His research focuses on political sociology and measurement in social science with current research including survey translation, internet data, and the detection of election fraud. He works on election observation missions as a statistical analyst for the Office for Democratic Institution and Human Rights and has worked on the BBC's election night coverage as a psephologist.

References

World values survey website, 2011.

Christian Bjornskov. How comparable are the gallup world poll life satisfaction data? *Journal of Happiness Studies*, 2010.

P. Bloom. *How children learn the meanings of words*. The MIT Press, 2002.

K.A. Bollen. Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1):605–634, 2002.

NH Carrier. A note on the measurement of digital preference in age recordings. *Journal of the Institute of Actuaries [JIA]*, 85:71–85, 1959.

R. Chaffin, R.K. Morris, and R.E. Seely. Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):225, 2001.

Stata Corporation. *Stata statistical software*. Stata Corp., 1999.

G. De Pauw, G.M. De Schryver, and P.W. Wagacha. A corpus-based survey of four electronic swahili english bilingual dictionaries. *Lexikos*, 19(0), 2009.

T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

Andrew Feldman. Spring cleaning for some of our apis, 2011.

J. Harkness. Questionnaire translation. *Cross-cultural survey methods*, 325: 35, 2003.

Janet Harkness, Ana Villar, and Brad Edwards. *Translation, Adaptation, and Design*. John Wiley and Sons, 2010.

A. Heath, S. Fisher, and S. Smith. The globalization of public opinion research. *Annu. Rev. Polit. Sci.*, 8:297–333, 2005.

R. Inglehart and C. Welzel. *Modernization, cultural change, and democracy: The human development sequence*. Cambridge Univ Pr, 2005.

T.P. Johnson. Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA Nachrichten Spezial*, 3:1–40, 1998.

Frank E Harrell Jr and with contributions from many other users. *Hmisc: Harrell Miscellaneous*, 2010. URL `http://CRAN.R-project.org/package=Hmisc`. R package version 3.8-3.

L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 39. Wiley Online Library, 1990.

G. King, C.J.L. Murray, J.A. Salomon, and A. Tandon. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(01):191–207, 2004.

Duncan Temple Lang. *RCurl: General network (HTTP/FTP/...) client interface for R*, 2010. URL `http://CRAN.R-project.org/package=RCurl`. R package version 1.5-0.1.

Duncan Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus.*, 2011. URL `http://CRAN.R-project.org/package=XML`. R package version 3.4-0.2.

Martin Maechler, Peter Rousseeuw, Anja Struyf, and Mia Hubert. Cluster analysis basics and extensions. 2005.

Adrian Mander. Plotbeta: Stata module to plot linear combinations of coefficients. Statistical Software Components, Boston College Department of Economics, January 2005. URL `http://ideas.repec.org/c/boc/bocode/s448702.html`.

J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176, 2011.

Microsoft. Bing translator api, 2011.

C.D. Nye, B.W. Roberts, G. Saucier, and X. Zhou. Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality*, 42(6):1524–1536, 2008.

E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

N. Pendar and S. Sharoff. Supervised lexical acquisition for persian from a web corpus. *Computational Approaches to Arabic Script-based Languages*, page 106, 2009.

E.O. Pérez. The origins and implications of language effects in multilingual surveys: A mimic approach with application to latino political attitudes. *Political Analysis*, 19(4):434–454, 2011.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. 3-900051-07-0.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Serge Sharoff. Leeds collection of internet corpora, 2011.

D. Stegmueller. Apples and oranges? the problem of equivalence in comparative research. *Political Analysis*, 19(4):471–487, 2011.

A. Thanopoulos, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of collocation extraction metrics. In *3rd International Conference on Language Resources and Evaluation*, volume 2, pages 620–625. Citeseer, 2002.

B. Tibi. *Arab nationalism: between Islam and the nation-state*. Palgrave Macmillan, 1997.

S. Verba. Cross-national survey research: The problem of credibility. *Comparative methods in Sociology: Essays on trends and applications*, pages 309–356, 1971.

Gregory R. Warnes, with contributions from Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley, Don MacQueen, Arni Magnusson, Jim Rogers, , et al. *gdata: Various R programming tools for data manipulation*, 2011. URL `http://CRAN.R-project.org/package=gdata`. R package version 2.8.2.

Gordon Willis, Martha Kudela, Kerry Levin, Alicia Norberg, Debra Stark, Barbara Forsyth, Pat Brick, David Berrigan, Frances Thompson, Deidre Lawrence, and Anne Hartman. *Evaluation of Multistep Survey Translation Process*. John Wiley and Sons, 2010.

A Computational Details

All the analysis of the meanings of words was performed in R [R Development Core Team, 2004]. The functions for retrieving the collocations from the Leeds Collection of Internet Corpora used the RCurl [Lang, 2010] and XML [Lang, 2011]. Manipulation of the text and data structures were handled using function from the base package [R Development Core Team, 2011], Harrel's miscellaneous package [Jr and with contributions from many other users., 2010], and the gdata package [Warnes et al., 2011].

The divisive hierarchical DIANA clustering algorithm is implemented in the cluster package [Maechler et al., 2005]. Additional checks on the dissimilarities data were performed using the Analyses of Phylogenetics and Evolution package [Paradis et al., 2004].

The analysis of the World Values Survey was performed using Stata [Corporation, 1999]. The coefficient graphs were created using the plotbeta package [Mander, 2005].

B Derivation of Relationship between Confounding Error Component and Individual Attributes

To see whether there is a violation of confounding error component's equivalence we look at a small subset of X_{ik}, Z_{ik}, and its effect on the errors. Since Z_{ik} is a subset of X_{ik}, shown in Equation 12, the relationships between X_{ik} and the error components also hold for Z_{ik}.[15]

[15] Z_{ik} will be referred to as a single variable here but the result also applies to a vector of confounding variables.

$$Z_{ik} \subset X_{ik} \tag{12}$$

Consequently the covariance of the true score and the potential confounding variable can be expressed in terms of the error components in Equation 13.

$$Cov(\hat{\theta}_{ik}, Z_{ik}) = Cov(\theta_{ik} + \varepsilon_{rik} + \varepsilon_{cik} + \varepsilon_{uik}\varepsilon_{sik}, Z_{ik}) \tag{13}$$

Given the covariance of the random, uniform and specification components with X_{ik} from Equations 3, 5, and 6 respectively, this can be simplified to Equation 14.

$$Cov(\hat{\theta}_{ik}, Z_{ik}) = Cov(\theta_{ik} + \varepsilon_{cik}, Z_{ik}) \tag{14}$$

In addition we assume that the true relationship is equivalent across countries in Equation 15.

$$\forall k : Cov(\theta_{ik}, Z_{ik} = e) \tag{15}$$

Combining this assumption with the relationship between the measured score and confounder derived in Equation 14, it is simple to show that the confounding error component's equivalence condition from Equation 8, that the covariance of ε_{cik} and X_{ik} is constant across countries, cannot hold.

C Number of Observations used in Logistic Regressions

Language cluster	Tradition important	Fight for your country	Proud of nationality	Importance of God
Western Europe	21,808	71,036	81,681	84,936
Chinese	1,664	4,701	4,933	2,514
Russian/Polish	3,835	13,072	15,036	13,023
Arabic	5,345	7,175	14,184	14,223

Table 7: Number of observations for used in each logit analysis by language cluster

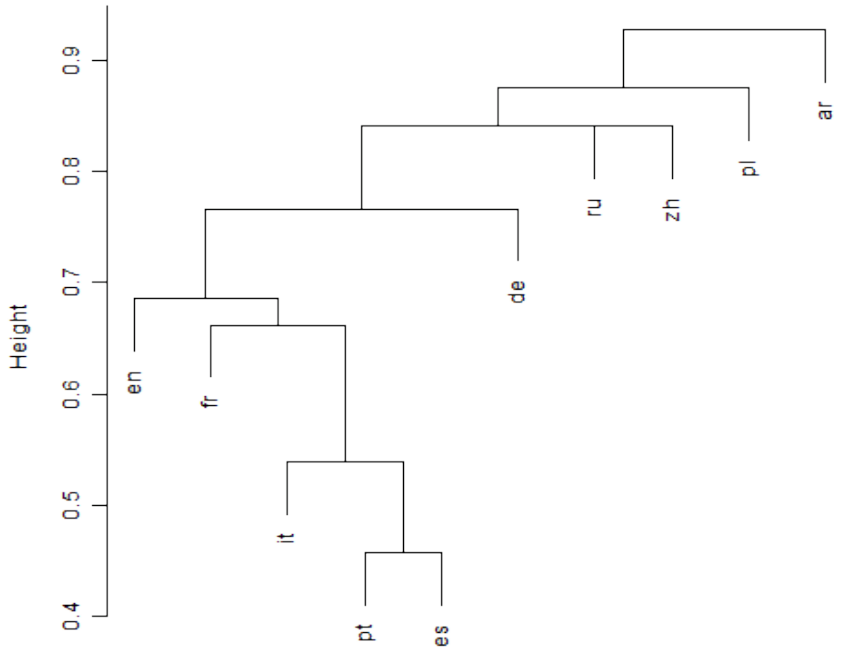


Figure 2: Hierarchical clustering dendogram of *economy* dissimilarities

D Linguistic Analysis of Other Words in Post-Materialism Question

4.1 A Stable Economy

Before the effects of the differing meanings of *ideas* can be predicted it is necessary to look at the translation of other important words in the same question. The word *economy* showed little evidence of clustering as can be seen in Figure 2. The languages break off at regular intervals without clear groups of languages. The collocations in the possible cluster of English, French, Italian, Portuguese and Spanish showed no difference in the types of collocations from the rest of the languages.

4.2 The Fight Against Crime

The translations of *fight* showed differences in meaning across languages and some evidence of clustering in Figure 3 with Italian and Spanish being particularly close as are English and Chinese. Examination of the shared collocations reveals that in Chinese and English a metaphor with violence is more prominent whereas all the other languages use terms that are more

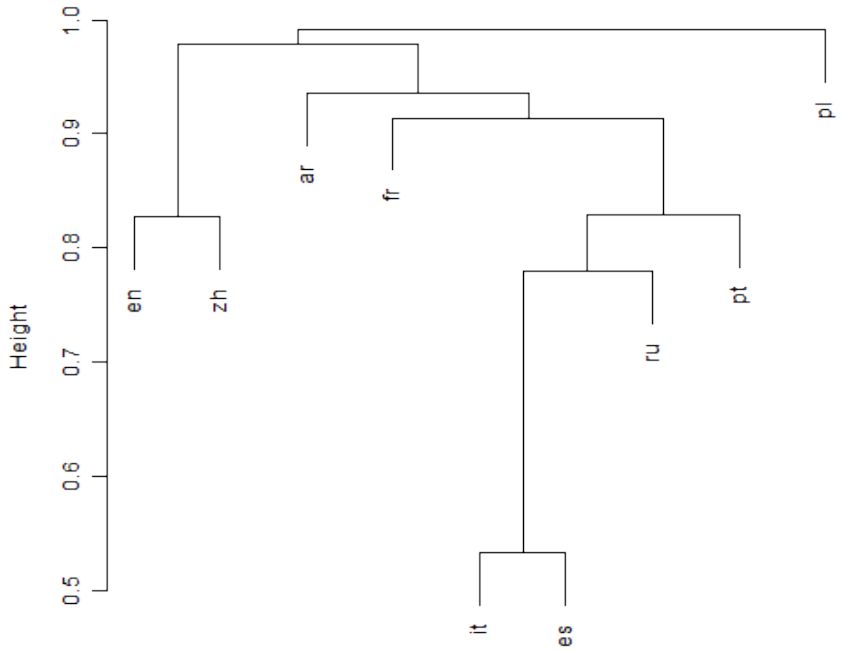


Figure 3: Hierarchical clustering dendogram of *fight* dissimilarities

general ways of reducing something although still implying . In other words the option "the fight against crime" refers to the same policy stance in every language (reducing crime) but the metaphor of combat is absent from the majority of translations. Although the exact meanings of the word *fight* may differ, it is not clear that it should greatly affect responses to the question as it is clear that the same type of policies are being referred to in all languages.

The word *crime* showed no important differences between languages. All languages showed associations with unlawful behaviour and responses to it such as police and prison.

4.3 Progress Toward a Less Impersonal and More Humane Society

he other post-materialist option in the question has several words that are important to the interpretation of the question: *progress*, *impersonal*, and *humane*.

The word *progress* shows some evidence of clustering in the dendogram in Figure 4. However, this clustering is shown not to be semantically important when we look at the collocations shared within each cluster in Table 4.3. The shared collocations between the cluster of English and Chinese do not differ

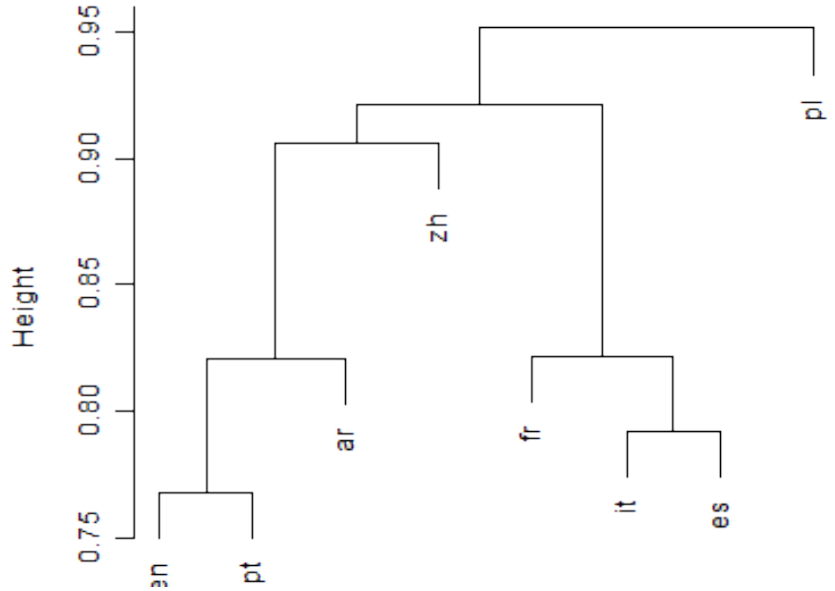


Figure 4: Hierarchical clustering dendogram of *progress* dissimilarities

greatly in meaning from those shared between the cluster of French, Italian, Portuguese, and Spanish. In both cases the primary meaning is related to development and movement. This meaning is also reflected in the shared collocations of the less clustered languages of Arabic and Polish which also share many of the same collocations as the other two clusters. In this case the clustering that is found does not appear to be reflecting something that will be problematic for the survey question.

It was not possible to run the analysis on the translations of *humane* in German, Polish, and Russian because it didn't appear often enough in their corpora. The DIANA analysis showed that there were two clusters of meaning. The first consists just of Arabic and English. The second cluster contains French, Portuguese, Italian and Spanish.

It was not completely clear whether there was a difference in meaning between the clusters from the lists of shared collocations alone. The second cluster includes a wide range of quite philosophical words such as *existence*, *natural*, and *conciousness*. Arabic and English both include much more practical words on their lists such as *treatment* and *organization*. Indeed Arabic's list of words mostly refer to humanitarian situations whilst English's most commonly refer to animal cruelty with some mention of societal issues such as *euthanasia*. The main difference between the clusters is the generality of the term *humane* which is much more specific in English and Arabic.

English and Chinese		Portuguese, Spanish, French, and Italian		Arabic and Polish	
collocations	number of languages shared in	collocations	number of languages shared in	collocations	number of languages shared in
all	2	all	4	any	2
also	2	also	4	each	2
area	2	by	4	great	2
country	2	do	4	human	2
development	2	duty	4	may	2
economic	2	have	4	not	2
from	2	if	4	on	2
further	2	more	4	political	2
make	2	new	4	social	2
more	2	not	4	society	2
need	2	power	4	what	2
new	2	without	4	which	2
on	2	allow	3	world	2
social	2	development	3		
there	2	go	3		
will	2	knowledge	3		
work	2	life	3		
year	2	see	3		
		social	3		
		society	3		
		want	3		
		well	3		
		world	3		

Table 8: Collocations shared within *progress* clusters

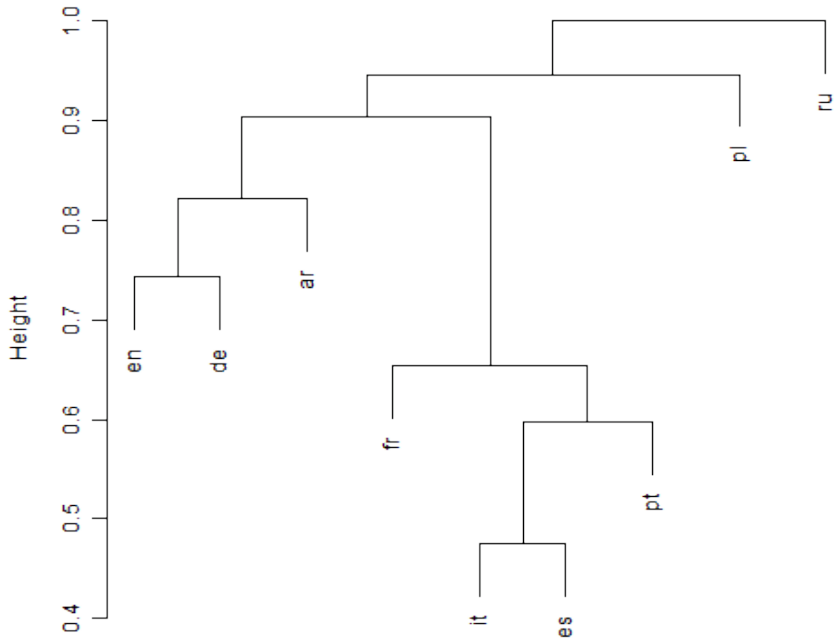


Figure 5: Hierarchical clustering dendogram of *humane* dissimilarities

Although this difference in meaning is present, it is not clear what the effect of it should be. With the word *impersonal* in the question as well, it is unlikely that English speaking respondents would think primarily of animal cruelty in answering the question.

No language has more than 800 instances of *impersonal* in its corpus which was not sufficient to perform a reliable cluster analysis. A qualitative comparison of the collocations did not suggest that there were widely differing meanings across languages. The main themes that came up were a lack of emotion and objectivity. These themes highlight a potential ambiguity in the question as respondents might respondent differently depending on which meaning is salient. However, the addition of the word *humane* in the question probably reduces this risk as it signals a negative, connotation of *impersonal*.

4.4 Summary of Linguistic Analysis

The analysis of the keywords within the question do not clearly show any differences between the languages that would be problematic for interpretation with the exception of *ideas*. However the difficulty in running the full analysis due to low frequency of the words within the corpora suggest that the question may be difficult to interpret in some languages as the vocab-

ulary will be unfamiliar to some respondents. In particular it is worrying that the post-materialist responses use more complicated vocabulary than the materialist options as this could introduce a bias towards these responses among less educated respondents. Since the major thesis of post-materialism is that economic development increases post-materialist attitudes [Inglehart and Welzel, 2005] this bias could be problematic for these findings. A plausible alternative interpretation would be that development increases education which leads to a greater proportion of respondents who are familiar with the vocabulary used in the post-materialist options in the survey.