UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Ana Slavec

Izboljševanje ubeseditve anketnih vprašanj z jezikovnimi viri

Improving survey question wording using language resources

Doktorska disertacija

Ljubljana, 2016

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Ana Slavec

Mentor: prof. dr. Vasja Vehovar

Somentor: prof. dr. Jon Krosnick

Izboljševanje ubeseditve anketnih vprašanj z jezikovnimi viri

Improving survey question wording using language resources

Doktorska disertacija

Ljubljana, 2016

Acknowledgments

First, I would like to thank my supervisor, **Vasja Vehovar**, for his trust and continuous support of my PhD study; that went beyond what was expected. He closely monitored the progress of this dissertation and readily provided guidance. One of the first lessons I learned from him is that science is not done alone; it is collaborative. He included me in several projects and endorsed my participation at scientific meetings, which was important for my professional development.

I want to thank my co-supervisor, **Jon Krosnick**, for the interest and enthusiasm he has shown for my research. Discussions with him and his unpublished book chapter on question wording helped me reflect on the initial overly ambitious proposal and narrow it down to a manageable scope. Moreover, his comments on my results were very thoughtful and concise.

I am also grateful to the two other committee members, **Darja Fišer** and **Valentina Hlebec**, who provided valuable feedback from their respective disciplines. I appreciate all four committee members reading and commenting on my manuscript, particularly given the short time frame.

In addition to my committee, there are several other individuals who helped me on this journey. I particularly want to thank **Simon Krek** and **Nataša Logar Berginc**, who, along with Darja Fišer, introduced me to the field of corpus linguistics and semantics. This was new to me, and I learned much in the process.

In this project, we developed an application that would support my research work. I thank **Mojca Mikac**, **Marko Grobelnik** and the others at the Artificial Intelligence Laboratory at the Jožef Stefan Institute who were involved in the development of the prototype. In addition, I want to thank **Gorazd Veselič**, who integrated it in the 1KA web survey tool.

I also want to acknowledge **Timo Lenzner**, whose work on the psycholinguistic determinants of question difficulty inspired my experiments, for providing feedback on the research design of the pilot experiment that was presented in Chapter 3. The results were also published in a special issue of the journal *Psihologija* about the psychology of survey participation and response. I thank the editor, **Michael Bosnjak**, and the two anonymous reviewers for their helpful comments, which improved the paper.

My project involved consultations with several survey methods experts who reviewed a set of survey questions. I am grateful to the **81 experts** who participated in the study that was presented in Chapter 4. Their reviews were of central importance for my research.

I would particularly like to thank the COST Webdatanet Action (IS1004) network, where I led the Evaluation of Questionnaire Quality task force. Thank you to everyone who attended the task force meetings, especially **Christopher Antoun** and **Melanie Revilla**, who provided valuable comments on the design of my experiments. At one network event, I met **Ray Poynter**, who pointed out the existence of false friends, such as the word 'sympathetic', which I included in my studies. Webdatanet supported my visit to the Amsterdam Institute for Advanced Labour Studies (AIAS), where I worked on the evaluations of the Wageindicator questionnaire presented in Chapter 4. I thank **Kea Tijdens** and the others at AIAS for their hospitality.

For Chapter 5, I must thank **Anja Mohorko** and **Gordon Willis**, who provided helpful feedback on the design of the cognitive interviews experiment. Anja also cognitively pre-tested the questionnaires for the international exchange students in the preliminary study. In addition, I must thank **Dominik Bašelj** for helping me code the open-ended answers in the cognitive interviews and for helping pre-test the questionnaires.

For Chapter 6, I thank **Mario Callegaro** for introducing me to **Sarah Cho** and **Jon Cohen** from Survey Monkey. Together, we discussed my main experiment that was carried out at their Audience Panel. Many thanks to the Audience team for their support.

A special thank you goes to my colleagues at the Centre for Social Informatics at the Faculty of Social Sciences at the University of Ljubljana. In particular, I thank **Anže Sendelbah**, who helped develop the questionnaires used for the split-ballot experiments and read several drafts of the individual chapters of my dissertation. He, **Nejc Berzelak** and **Andraž Petrovčič** also provided methodological advice on what statistical techniques to use. I thank **Barbara Brečko** and **Miha Matjašič**, who read and helped edit parts of my dissertation, and I thank **Tina Dolenc** and **Marjana Vrh**, who helped me with various administrative duties.

Finally, I want to thank my family and friends for their patience and acceptance, even though I barely had time for them in the last months and years.

Koper, 15th of June 2016

Izboljševanje ubeseditve anketnih vprašanj z jezikovnimi viri

Povzetek

Besede, ki jih pogosto uporabljamo v vsakdanjem govoru, prepoznamo in obdelamo hitreje kot besede, ki jih uporabljamo manj pogosto. Zato je v situacijah, kot je oblikovanje anketnih vprašanj, ko je povečevanje razumljivost besedila osrednjega pomena, zaželena uporaba običajnejših izrazov. Nepoznani izrazi so bili v literaturi namreč izpostavljeni kot ena od značilnosti besedila, ki vpliva na njegovo razumljivost. Kot so pokazale pretekle raziskave, lahko težave z razumljivostjo pomenijo povečano breme za anketiranca, daljši čas odgovorov, več neodgovorov spremenljivke, več prekinitev odgovarjanja in druge neželene vzorce odgovarjanja, ki vodijo do nižje kakovosti odgovorov. V določeni meri lahko težave z razumljivostjo anketnih vprašanj zaznamo z metodami predtestiranja in evalvacije vprašalnikov, kot so kognitivni intervjuji in ekspertne ocene. Obe metodi lahko potencialno napovesta problematična vprašanja, vendar so manj poznani izrazi specifičen problem, ki ga ni enostavno zaznati. Poleg tega sta omenjeni kvalitativni metodi zelo zahtevni z vidika porabljenega časa in drugih resursov.

V disertaciji predlagamo nov pristop, ki temelji na jezikovnih virih, kot so besedilne korpusi in leksikalne baze, ki bi lahko razvijalcem anketnih vprašalnikov služil kot dopolnilo tradicionalnim metodam evalvacije vprašanj. Besedilni korpusi so namreč velike zbirke besedil v naravnem okolju, ki se lahko uporabijo kot mera (ne)poznanosti določenega izraza. Višja je frekvenca v korpusu, bolj je beseda poznana splošni populaciji. Poleg tega lahko uporabimo še leksikalne baze, in sicer kot vir sopomenk in drugih alternativnih ubeseditev, s katerimi lahko potencialno problematične besede zamenjamo s pogostejšimi alternativami, po možnosti z enakim pomenom.

V empiričnem delu disertacije pristop, ki temelji na jezikovnih virih, uporabimo na treh študijah primera z različnimi vprašalniki, temami in vzorčnimi populacijami. Na podlagi besedilnih korpusov in leksikalnih baz razvijemo visoko frekventne in nizko frekventne različice istih vprašalnikov. Te različice nato evalviramo z ekspertnimi ocenami, kognitivnimi intervjuji in eksperimenti z deljenim vzorcem na vzorčni populaciji.

Najprej smo izvedli preliminarno pilotno študijo na dveh anketnih vprašalnikih za študente na mednarodni izmenjavi na Univerzi v Ljubljani, pri čemer je bil eden v angleškem (za prihajajoče študente) in eden v slovenskem jeziku (za odhajajoče študente). Oba vprašalnika smo evalvirali z jezikovnimi viri in razvili dve različici obeh vprašalnikov, eno z nizkimi frekvencami ubeseditev in drugo z visokimi frekvencami ubeseditev. Skupaj sta se obe angleški različici razlikovali v 23 ubeseditvah, slovenski pa v približno 40 ubeseditvah. Obe različici smo nato primerjali v dveh eksperimentih z deljenima vzorcema, kjer je bila polovica vzorca naključno dodeljena kontrolni skupini, ki je odgovarjala na različico z nizkimi frekvencami, in drugi polovici, ki je bila dodeljena eksperimentalni skupini, ki je odgovarjala na različico z visokimi frekvencami. Rezultati so pokazali, da je bilo manj prekinitev odgovarjanja v dveh različicah z visokimi frekvencami. Poleg tega so anketiranci v slovenski različici z višjimi frekvencami alternativnih ubeseditev poročali o nižjem številu manj razumljivih besed. Čeprav je imela pilotna študija vrsto omejitev, pa je dobro osvetlila smeri raziskovanja v osrednjem empiričnem delu disertacije.

Druga empirična študija je primerjala pristop na podlagi besedilnih korpusov z ekspertnimi ocenami za zaznavanje nepoznanih izrazov. Dva niza anketnih vprašanj sta bila izbrana kot študiji primera: prva je bila izbor osmih anketnih vprašanj (sedem različnih ubeseditev) iz vprašalnika WageIndicator o plačah in delovnih pogojih, druga pa je bila izbor osmih vprašanj (12 postavk in 12 različnih ubeseditev) iz baze anketnih vprašanj PEW. Oba vprašalnika smo evalvirali na podlagi jezikovnih korpusov, alternativne izraze pa smo poiskali v leksikalni bazi WordNet; za vsako postavko smo izbrali nekaj besed, ki so jih potem evalvirali eksperti. Eksperte smo prosili, naj ocenijo primernost različnih ubeseditev, označijo, katere bi izbrali, in komentirajo svoje odgovore. Skupaj je sodelovalo 81 globalnih ekspertov s področja anketne metodologije. Rezultati so pokazali, da se evalvacije ekspertov in besedilni korpusi ujemajo za več kot polovico postavk, v večini ostalih postavk pa tudi ni bilo izrazitejših razlik. Večje razlike so se pojavile le v nekaj primerih, kar lahko večinoma pojasnimo s tem, da besede niso imele povsem enakega pomena in zato v konkretnem kontekstu niso zamenljive. Z drugimi besedami, alternativne ubeseditve niso bile enakovredni sinonimi. Kljub temu smo lahko zaključili, da lahko opisani polavtomatski pristop na podlagi korpusov v znatni meri nadomesti zahtevne (v smislu porabe časa in resursov) ekspertne evalvacije.

Tretjo empirično študijo sestavlja 122 spletnih kognitivnih intervjujev, kjer smo udeležence vprašali bodisi po definiciji določene ubeseditve v anketnem vprašanju bodisi po njenem parafraziranju. V celoti smo evalvirali 13 postavk, vse iz zgoraj omenjenega niza vprašanj PEW. Udeležence smo rekrutirali z uporabo globalne platforme Prolific Academic za množično sodelovanje (»crowdsourcing«). Študija je bila osnovana na eksperimentu z deljenim vzorcem, saj je bila polovica sodelujočih naključno razvrščena v različico z izvirnimi vprašanji PEW, polovica pa v različico z izboljšanimi (sedem primerov) ali s poslabšanimi (šest primerov) vprašanji. Ugotovili smo, da v primeru, ko uporabimo nizko frekventno besedo, to besedo udeleženci praviloma definirajo oziroma parafrazirajo z njeno bolj frekventno alternativo. V primeru bolj frekventnih ubeseditev smo skupno našteli tudi višje število različnih definicij in parafraz v primerjavi z njihovimi nizko frekventnimi alternativami. V nekaterih primerih smo to pojasnili z višjim številom pomenov (v bazi WordNet), kar nakazuje na problem večje dvoumnosti teh izrazov. Poleg tega smo ugotovili tudi določene razlike med tistimi, ki jim je angleščina materni jezik, in ostalimi.

Četrta in glavna empirična študija je bila eksperiment z deljenim vzorcem, kjer smo primerjali štiri različice istega vprašalnika PEW: izvorno, izboljšano (11 zamenjav z bolj frekventnimi ubeseditvami), slabšo (16 zamenjav z manj frekventnimi ubeseditvami) in najslabšo (34 zamenjav z izrazito manj frekventnimi ubeseditvami). Eksperiment je potrdil, da poznanost izraza, kot jo merimo s frekvencami v korpusih, lahko vpliva na različne vidike kakovosti anketnih podatkov, zlasti na prekinitve odgovarjanja in subjektivne ocene težavnosti odgovarjanja. Zaznali smo tudi daljši čas odgovarjanja in za nekatere postavke tudi več odgovorov »ne vem« ter večjo težnjo k strinjanju z odgovori. Vendar so bili učinki pri zmernem (izboljšana in slabša različica) variiranju alternativnih ubeseditev večinoma majhni. Videti je, da manjše število zmernih (v smislu povečane ali zmanjšanje frekventnosti) alternativnih ubeseditev ne povzroči izrazitejših sprememb pri večini indikatorjev kakovosti odgovarjanja. Večji učinki pa se pokažejo pri seštevanju izrazitejših sprememb.

Rezultati so potrdili, da je na osnovi besedilnih korpusov, leksikalnih baz in slovarjev možno razviti postopek, na podlagi katerega lahko učinkovito zaznavamo problematične

ubeseditve anketnih vprašanj in na tej osnovi predlagamo tudi alternative ubeseditve. Poleg tega rezultati kažejo, da je v večini primerov pristop na osnovi korpusov primerljiv z ekspertnimi ocenami in kognitivnimi intervjuji. Vendar je pomembno, da pri tem upoštevamo specifičnost zasnove različnih korpusov in se ne omejimo na evalvacijo le posameznih besed, ampak preverimo tudi daljše besedilne nize.

V prihodnosti je treba ta pristop še nadalje empirično evalvirati, zlasti v smeri iskanja kritičnega nivoja sprememb ubeseditev (osnovanih na korpusnih frekvencah med alternativnimi sinonimi v nizu), ki lahko ogrozijo kakovost anketnih podatkov. Poleg tega so za odkrivanje ključnih faktorjev potrebne še sistematične metaanalitične študije raznih sekundarnih podatkov. Smiselno pa je razvijati tudi potenciale, ki jih ima opisani pristop za vključitev v programska orodja za spletno anketiranje.

Ključne besede: ubeseditev vprašanja, nepoznani izrazi, jezikovni viri, metode za evalvacijo vprašalnikov, eksperiment z deljenim vzorcem.

Improving survey question wording using language resources

Abstract

Words commonly used in daily speech are recognised and processed more quickly than words that are less commonly used. Thus, the use of more common words is preferred in contexts where maximising text comprehensibility is of central importance, which is usually also the case in survey questions. In fact, unfamiliar words have often been indicated in the literature as one of the text features that can affect question comprehensibility. As previous studies have shown, comprehensibility issues might lead to an increase in response burden, longer response times, more item non-response and drop-outs, and other undesired respondent behaviour that can decrease response quality. To a certain extent, comprehensibility problems in survey questions can be detected with pre-testing and evaluation methods, such as cognitive interviews and expert reviews. Both have been shown to have the potential to positively predict problematic questions; however, unfamiliar words are a specific problem that might not be detected easily. Moreover, these qualitative methods are very demanding in terms of both time and resources.

In this dissertation, a new approach based on language resources, including text corpora and lexical databases, is proposed to assist questionnaire designers as a supplement to traditional question evaluation methods. Text corpora are large samples of language in natural contexts that can be used as estimates of wording unfamiliarity. The higher the frequency in corpora, the more familiar a word is to the general population. In addition, lexical databases are used as a source of word synonyms and other alternative wordings that can replace a potentially problematic word with a higher frequency wording, preferably with the same meaning.

In the empirical part of this dissertation, the linguistic resources approach is applied to three case studies with different questionnaires, topics and sample populations. Based on linguistic corpora and lexical databases, we develop low-frequency and high-frequency versions of the same questions. We then evaluate the different versions using expert reviews, cognitive interviews, and split-ballot studies on the sample population.

First, a preliminary pilot study on two web survey questionnaires for international exchange students at the University of Ljubljana was conducted, one in the English language (for incoming students) and the other in the Slovenian language (for outgoing students). The two questionnaires were evaluated with linguistic resources and two versions were developed for each, one with low-frequency wordings and one with their high-frequency synonymous wordings. In total, the two English versions differed in 23 wordings and the two Slovenian versions in about 40 wordings. The versions were then compared in two split-ballot experiments, where half of the sample was randomly assigned to the control group that responded to the low-frequency version and the other half were assigned to the experimental group that responded to the improved version with more frequent wordings. The results show that there was a lower drop-out in the two experimental versions. In addition, respondents to the improved Slovenian version reported a lower number of words that were not understood. Despite various limitations of this pilot study it successfully traced the directions for main empirical studies.

The second empirical study involved comparing the text corpora approach with expert reviews to detect unfamiliar wordings. Two sets of survey questions were selected as case studies: one was a selection of eight survey questions (seven different wordings) from the WageIndicator questionnaire on wages and working conditions, and the second was a selection of eight questions (12 items and 12 different wordings) from the PEW database of survey questions. Both questionnaires were evaluated using text corpora and alternative wordings were searched for in the WordNet lexical database; so for each item, we selected a set of alternative wordings to be evaluated by experts. Experts were asked to evaluate the appropriateness of different wordings, indicate which they would choose, and comment on their responses. In total, 81 experts participated. The results show that for more than half of the items, there was a full match between evaluations based on text corpora and those provided by experts; while for the majority of the remaining items, there were only minor differences. In a few cases, there were larger discrepancies, the main reason for which was that these words did not have the same meaning and were not interchangeable in that context – that is, they were not synonyms. Thus, the semi-automated corpora approach can replace resource-demanding expert evaluations.

The third empirical study consisted of 122 web-based cognitive interviews, where participants were asked to either define a certain wording used in a survey question or to paraphrase the full question. In total, 13 items were evaluated, all from the abovementioned PEW items. Participants were recruited using the Prolific Academic crowdsourcing platform. This study was also based on a split-ballot experiment, as a random half of the participants evaluated original PEW items, while the other half responded to either improved (in seven cases) or worsened (in six cases) items. We found, as expected, that when presented with a low-frequency wording, respondents generally used its high-frequency alternative. Another finding was that there were a greater number of different definitions and paraphrases listed for high-frequency wordings compared to their low-frequency counterparts. In some cases, this was explained with a higher number of senses (i.e., meanings in WordNet), which is an indication of greater wording ambiguity. In addition, there were also some differences according to native and non-native speakers.

The fourth and main empirical study was a split-ballot experiment where four versions of the same PEW questionnaire were compared: original, improved (11 improved wordings), worse (16 worsened wordings) and the worst (34 worsened wordings). The experiment confirmed that the familiarity of question wordings measured with corpora frequencies can affect the response quality of survey data, particularly with respect to drop-out and subjective evaluations of response burden; furthermore, increased response times were observed, and for some items there were more ‘don’t know’ responses and acquiescence. However, the effects were mostly small. It seems that a small amount of changes does not produce much of a difference in most of the response quality indicators.

The results confirm that it is possible to develop a procedure based on text corpora, lexical databases and dictionaries that can effectively detect problematic question wordings and suggest alternatives. Moreover, the results show that in most cases the text corpora approach gives comparable results to expert reviews and cognitive interviews. However, it is important to take into account the specific design of different corpora and not limit only to the evaluation of single word frequencies, but the frequencies of strings of words.

In the future, the approach should be further empirically evaluated, particularly in the direction of finding the critical level of wording changes (based on corpora frequencies among alternative synonyms) that can damage the quality of the survey data. In addition, systematic meta-analytical studies of various secondary data shall be conducted to discover key factors in this complex matter. There is also the potential to incorporate this approach into survey questionnaire design tools.

Keywords: question wording, unfamiliar words, linguistic resources, questionnaire evaluation methods, split-ballot experiments

Contents

1 Introduction

1.1 Question wording and the questionnaire evaluation process

Surveys are the prevailing data collection method in quantitative social science research, marketing and official statistics. The development of the corresponding measurement instrument includes conceptualisation in the form of theoretical variables, operationalisation and the measurement of empirical variables. After their implementation, we can obtain observed values from respondents. No survey measurement is without errors, both random (variance) and systematic (bias) (Groves et al. 2009). Various sources for errors exist, and within this context high quality questionnaires are extremely important.

Writing good survey questions is a complex task where several decisions need to be made regarding conceptual and technical issues, such as question topic, question type and format, response categories, ordering of questions, visual aspects and question wording. The latter is perhaps the most difficult, as it relates to the selection and combination of words from several possibilities. The researcher who develops the questionnaire needs to be aware of the information requirements and be able to use the right words to formulate questions. Moreover, it is useful to have an understanding of the psychology of the response process and at least some methodological and statistical knowledge and familiarity with recommended practices. An additional advantage is having knowledge of available technology that can be used to improve questionnaire quality (Couper et al. 1998).

It is often not clear if question wording is an art or a science. Furthermore, although many textbook guidelines on question wording exist, researchers often rely only on common sense and experience when generating survey questions. However, question wording is a very complex characteristic as each question can be worded in numerous ways, and it is hard to estimate if different words and phrases are interchangeable – and if not, what the optimal wording would be. Given the difficulty and complexity of generating good questions, non-optimal wordings can easily occur. Even educated questionnaire developers have difficulties, because they use overcomplicated language

that is too demanding for the respondent (Sheatsley 1983, 200). Within this context, common words are preferred; for example, in their textbook on handcrafting a standardised questionnaire, Converse and Presser (1986) suggested using ‘main’ instead of ‘principal’, which is a less frequent word.

Research shows that respondents are quite sensitive to structural characteristics of questionnaires, while it is less clear to what extent they are sensitive to differences in question wording (Krosnick and Fabrigar, forthcoming). According to Krosnick and Fabrigar, a good question wording should, in theory, strive for univocality, meaning uniformity and economy of words. First, univocal wording means that the question is clearly focused only on the concept being measured and does not include other concepts (i.e., avoid prestige names, double-barrelled and leading questions). Second, it is uniform when the wording has a single meaning for all respondents; that is, when it avoids words with many possible interpretations, jargon, slang or colloquialisms, and abstract, ambiguous or emotionally charged words. Third, economy of words means that no more words than are needed to communicate an idea clearly should be used. On one hand, lengthier questions are more burdensome for respondents to process and interpret. On the other hand, it is more natural and easier to ensure uniform interpretations with more words. Some questions are more burdensome in an abbreviated form, while others are more so in a longer form, and there is no general rule to help us decide because it often depends on the context. This is also reflected in the most commonly mentioned guidelines in survey design textbooks. Moreover, some experimental studies have shown the importance of carefully selecting the most appropriate wording (e.g., Kalton et al. 1978; Duncan and Schuman 1980; Schuman and Presser 1981; Bradburn and Sudman 1983; Smith 1987; Rasinski 1989). Nevertheless, given the extensive range of possibilities for every concept, the area is still quite under-researched, and the sensitivity of respondents to differences in question wording is not yet completely clear.

The research on wording is related to broader research on cognitive aspects of survey methods (CASM) that draws on psychological theories of language comprehension, memory and judgment (e.g., Tourangeau 1984; Sirken et al. 1999; Tourangeau et al. 2000; Schwarz 2007). Following the ESCRIME (encoding, storage, comprehension, retrieval, integration, mapping, editing) stages in the response process, wording is very

important in the comprehension (clarity of definitions and instructions), retrieval (reinforcing reference periods, improving recall, motivation), and judgment and response stage (desensitise items) (Schaeffer and Dykema 2011). In the proposed dissertation, we further narrow the focus only to the comprehension stage, for which wording is of central importance.

In survey research literature, different authors use different typologies of comprehension problems related to wording, and they typically expose the following issues: ambiguity and conceptual variability, excessive complexity, vague concepts and quantifiers, unfamiliar terms, and false inferences (Tourangeau et al. 2000; Lenzner 2011, 2012). These kinds of problems can affect response quality in various ways and contribute to an increase in non-response and measurement error in survey data.

In this dissertation, we further narrow the focus to unfamiliar words. As cognitive psychology research has shown, words commonly used in daily speech are recognised and processed more quickly than less commonly used words, which is labelled the 'word frequency effect' (Howes and Solomon 1951; Broadbent 1967). One way to operationalise how frequent and familiar a certain word is within a language is to use text corpora. Text corpora are large samples of language in a natural context, such as books, newspapers, magazines and Internet resources, which are merged to generate structured databases according to specific criteria and aims. Text corpora can be used to generate wording frequency estimates, both for single words and strings of words. Words that are less frequent in corpora are supposedly less familiar to readers and decrease text comprehensibility. Correspondingly, eye-tracking studies have shown that the use of low-frequency words triggers longer gaze times (Inhoff and Reyner 1986; Jurafsky 2003). This also holds true for survey questions, as confirmed by eye-tracking studies (Lenzner et al. 2011) and with quantitative indicators of response quality (Lenzner 2012).

However, text corpora are not commonly used by questionnaire designers. Instead, they mostly rely on traditional question evaluation methods. On one hand, quantitative approaches, such as the abovementioned experimental studies, exist that compare different versions of the same questions in a **split-ballot** technique (Rugg 1941; Cantril 1944). However, as this requires fielding a survey, it is usually used in the final steps. On the other hand, pre-testing and evaluation methods exist that can be used to detect

problems in survey questions before conducting the pilot and/or final survey, such as **expert reviews** and **cognitive interviews** (Presser et al. 2004; Madans et al. 2011). Yet, these qualitative approaches, which are based on personal judgement, can have reliability issues and still be quite resource-consuming.

In addition, there are also evaluation procedures based on computerised models that do not require any additional data collection. In fact, modern information communication technologies (ICT) have revolutionised the survey process in recent decades, particularly by further integrating the entire process (Vehovar et al. 2014). Within this context, intelligent ICT support is also being extended to the questionnaire development stages. One important application is the **Survey Quality Predictor (SQP)**, which is based on a meta-analysis of multi-trait, multi-method experiments (MTMM) for more than 3,000 questions and allows users to obtain predictions of reliability and validity for any new question (Saris and Gallhofer 2007) for all European languages. However, this is also very time consuming, as more than 40 question characteristics have to be manually coded for each variable. Another shortcoming – within our context of wording problems – is that the method mainly focuses on structural and formative characteristics of survey questions and only on a few linguistic indicators (length of syllables, words and sentences in question introduction/request).

Here, the previously mentioned text corpora and other linguistic resources could be used to supplement current question evaluation methods. In fact, even though linguistic corpora and other resources are freely available for academic use and could be useful for computing indicators of word unfamiliarity to assist survey questionnaire design (Krosnick and Fabrigar, forthcoming), they remain underutilised in survey research. An exception is the **Question Understanding AID (QUAID)**, a computerised method for question evaluation in the English language which focuses on psycholinguistic determinants of question complexity (Graesser et al. 1999, 2000, 2006). It identifies five problems in survey questions: Low-frequency words; Vague or imprecise relative terms; Vague or ambiguous noun phrases; Complex syntax; and Complex logical structures. All five aspects were also considered in Lenzner's list of determinants of question comprehensibility (Lenzner 2010, 2012). However, our experience with the QUAID tool is that it gives a lot of false positives (confirmed also by Graesser et al. 2000).

Another shortcoming of the QUAID method is that it does not offer any suggestions of alternative wordings, synonyms or hyponyms that would be more familiar to the respondent and would improve the question. Researchers have to find them on their own (e.g., thesauri in word processors), which is a subjective and non-systematised task. On the other hand, the WordNet lexical database, which contains strings of inter-changeable synonymous words (synsets) and is considered to be a gold standard in computational linguistics, has not been utilised for retrieving synonymous words for the purposes of questionnaire design as of yet, at least to our knowledge.

In fact, in most of the abovementioned experimental studies, the effect of changes in wording on response distributions was studied under the assumption that the alternatives do not have the same meaning and respondents interpret them differently. In this dissertation, however, we mostly focus on situations where two or more words are interchangeable as they share the same meaning – in other words, they are synonyms. Synonymous wordings can be found in thesauri and semantic lexica such as WordNet, where the definition of this relationship is applied as follows: substituting the two words does not change their meaning in a certain context (Miller 1995). For instance, the adjectives ‘main’ and ‘principal’ mentioned at the beginning of this section have one meaning in common.

However, even in this case, the two words are not necessarily completely interchangeable, since different wording alternatives are not always equivalent in terms of familiarity and this can affect question comprehensibility. In addition, the levels of familiarity and comprehensibility can differ across different groups of respondents, depending on their education and other characteristics (Nation and Waring 1997).

Moreover, it should be noted that the use of text corpora and semantic lexica are not the only linguistic approaches that could be useful in improving survey question wording. In fact, readability metrics such as the Flesh-Kincaid readability test (Flesch 1943; Kincaid et al. 1975) provide a score of comprehension difficulty. Yet, these measures were developed for longer texts and likely do not work optimally on very short texts such as survey questions. In any case, as stated earlier in this chapter, in this dissertation we narrow the focus to only unfamiliar terms.

1.2 The research problem

Even though a fairly large body of literature and experimental work has been devoted to questionnaire design, several issues remain understudied, particularly in relation to the potential use of language resources. While structural characteristics of questions and questionnaires (e.g., question and answer type, number and label of categories, question order) have been comprehensively examined in the literature, it is still not completely clear how variations of question wordings affect data quality. Most of the research on question wording is decades old and there are only a few recent examples.

In particular, given the above overview of the relationship between language resources and the questionnaire development process, there is obviously a gap in knowledge regarding linguistic properties of survey questions, such as frequency in linguistic corpora and corresponding question ambiguity. In addition, subtle problems in the question comprehension stage exist that qualitative pre-testing methods, such as cognitive interviews and expert reviews, might not be able to detect (Graesser et al. 1999). Although certain applications like QUAID and SQP (mentioned in Section 1.1) have been designed to help researchers detect some problems from the linguistic perspective, these tools are not used often by survey researchers – possibly due to their shortcomings, as described in Section 1.1. In addition, they also leave many issues related to the utilisation of language resources in questionnaire development unaddressed.

Correspondingly, there is a lot of room for improvement in terms of how language resources such as text corpora and WordNet are applied in the survey questionnaire development process, particularly in pre-testing. These potentials are also the main focus of this dissertation. More precisely, we address issues related to components and characteristics of language resources which could be useful for survey designers. We are particularly interested in comparisons of approaches based on language resources with standard pre-testing methods as well as on studying how a language resources-based approach can help design question wordings that are less burdensome for respondents, which can in turn improve data quality.

1.3 Main thesis and research questions

Based on the above-defined research problem, the aim of this dissertation is to develop and evaluate a new methodological approach for the evaluation of question wording based on computational linguistic resources which can potentially supplement other questionnaire pre-testing methods. Correspondingly, our main thesis is that it is possible to **develop a procedure based on text corpora and lexical databases that can effectively detect problematic question wordings and suggest alternatives**.

On one hand, this thesis builds on ideas from Krosnick and Fabrigar's (forthcoming) chapter on question wording, including the notion that the complexity of the respondent task could potentially be alleviated by using specialised computer programs that would detect and highlight words which are ambiguous, unfamiliar, abstract or complex, and would also suggest synonyms which are clearer, more familiar, specific and simple. On the other hand, it updates the work of Graesser (2006) and Lenzner (2011) with respect to making the procedures for detecting comprehensibility problems semi-automated.

The main research idea is to theoretically elaborate an approach based on language resources and outline the corresponding operational procedures. In addition, the goal is also to conduct thorough empirical examinations and comparisons with standard pre-testing methods to demonstrate that the corresponding procedures can be effectively used to significantly improve the quality of a survey questionnaire – that is, by becoming a useful addition (or even a replacement) to other survey questionnaire pre-testing methods. In addition to the questionnaire pre-testing and evaluation approach, one additional goal is to preliminarily check how this approach could be used more actively, in the sense of offering suggestions for wording improvements during the questionnaire development process itself.

Based on the problem identified above, the main research idea, and the goals of our research, we can formulate the following key research questions:

1. How can linguistic resources be selected and combined so that questionnaire developers can detect low-frequency wordings in survey questions?
2. Do experts consider wording alternatives with higher frequencies found by using linguistic resources more appropriate to use compared to low-frequency wordings with the same meaning?

3. Do participants in cognitive interviews demonstrate a better understanding of wordings with higher frequencies than wordings with lower frequencies?
4. Is the response quality in surveys that use wording alternatives with higher frequencies better than in surveys that use lower frequency wordings?
5. How can the consideration of alternative wordings with higher frequencies be integrated into the process of questionnaire development?

The above research questions also determine the structure of this dissertation, which we outline in detail in the next section.

1.4 Thesis structure

Following the above elaborations, in the proceeding chapters we first address the theoretical background related to the potential of using language resources in the questionnaire development process (Chapter 2). Specifically, we examine the relationship between word familiarities and the wording frequencies effect, where we provide an overview of the potentials of language resources. We also review past research related to the implementation of text corpora approaches in survey questionnaire development. Finally, we introduce two standard evaluation methods, expert evaluations and cognitive interviews, which will be used later in this dissertation.

Next, in Chapter 3, we begin with a preliminary pilot study conducted with two small groups of exchange students at the University of Ljubljana: Slovenian students who participated in an exchange study abroad, and foreign students who were in Ljubljana for their student exchange in the previous year. The questionnaire was written in the Slovenian language for the first group and in English for the second group; however, the questions contained the same content. We evaluated the two survey instruments (questionnaires) using text corpora; and in one version, the problematic wordings were replaced with more familiar wording alternatives. The two Slovenian versions were also pre-tested using cognitive interviews analysed as part of another dissertation, and it was found that in the complex version, participants were less likely to report other kinds of errors (Mohorko 2015). The four versions were then administered to students in a split-ballot experiment: a standard method for questionnaire wording evaluations. We observed changes in item non-response, drop-out, response times, satisficing, and

subjective evaluations of cognitive difficulty, which we measured at the end of the questionnaire.

This preliminary pilot study already implemented the main idea of this dissertation; however, it had serious limitations: it used a very specific population, an ad-hoc questionnaire and a very small sample. In addition, it did not use any expert evaluations and only a limited set of cognitive interviews to compare them with the text corpora approach. Nevertheless, the experience obtained from the pilot study helped in designing the main empirical part of the research, where a rigorous approach was first used to select the question items and the corresponding cases where wordings would be changed. With the same level of rigor, the cases were then evaluated by experts and via cognitive interviews. In addition, a careful research experiment was designed and implemented on a large sample from one of the leading US online panels (Survey Monkey Audience).

In Chapter 4, we discuss the use language resources to carefully select the question items from the set of questions related to PEW (a non-profit, non-partisan US think tank that provides information about public issues) survey research on terrorism. We also selected the question items from the set of questions in the WageIndicator (a global web survey whose mission is to increase labour market transparency). In the next step, we used language resources to carefully select – based on differences in potential wording frequencies – the wording cases, which were then exposed to expert evaluations. Finally, around 50 experts evaluated the alternatives in both case studies, and we compared their results with corpora frequencies.

In Chapter 5, we describe how cognitive interviews were used to evaluate a selection of 13 question items from the PEW questionnaire; for each of them, two wording alternatives were tested, one with a lower wording frequency and one with a higher wording frequency. The crowdsourcing platform Prolific Academic was used to recruit 120 respondents. The results of online cognitive interviews were then compared with corpora frequencies.

In Chapter 6, we present the central empirical research study, which implemented the PEW questions in an online survey using the Survey Monkey Audience online panel. We prepared four versions of the PEW questionnaire (evaluated in Chapter 4 and

Chapter 5) and field-tested it with a **split-ballot experiment** for selected alternative wordings. Besides the original version, one version was improved (i.e., wordings with higher frequencies were used), one was moderately worsened (i.e., wordings with lower frequencies were used), and one was radically worsened (i.e., more than twice the number of wording changes compared to the moderately worse version). Altogether, 16 question items were subject to variation in 38 wording cases. Due to various specific wording alternatives (words or strings of words), which we will discuss in detail, 81 different wordings were used in total. For alternative wordings, we observed differences in response quality for the following aspects: response times, drop-out rates, 'don't know' answers, acquiescence, and subjective respondent estimates of cognitive difficulty and wording unfamiliarity. Furthermore, for the main indicators, we also controlled for the effect of education, native language, and gender. Finally, we summarised the results of the experiment and compared it to text corpora frequencies.

In addition, we also designed a split-ballot experiment for the WageIndicator study, but there was some delay in the start of the experiment, which was planned to begin in early 2015 but was postponed until February 2016. The data are still being collected and as such were not included in this dissertation.

To sum up, in Chapters 3–6, we elaborate on three case studies (questionnaires) that we evaluated with four different methods: corpora frequencies, expert reviews, cognitive interviews and a split-ballot quantitative study, as illustrated in Table 1.1.

Table 1.1: Overview of the characteristics and appearance of the three empirical studies

	Language	**Corpora Frequencies**	**Expert Review**	**Cognitive Interviews**	**Split-ballot experiment**
Case Study 1: International Exchange Students	English; Slovenian	February and March 2014 **(Chapter 3)**	x	March 2014 (Mohorko 2015)	April-May 2014 (**Chapter 3**)
Case Study 2: WageIndicator	English; (Slovenian)	February and March 2015 **(Chapter 4)**	May to Sept . 2015 **(Chapter 4)**	x	(February 2016-2017)
Case Study 3: PEW questions	English	April and May 2015 (**Chapter 4)**	June to Sept. 2015 **(Chapter 4**)	September 2015 **(Chapter 5)**	October 2015 **(Chapter 6)**

In Chapter 7, we summarise and discuss the main findings from the pilot study and the main study, as well as insights gained from cognitive interviews and expert evaluations. In addition, we also present the prototype application that we developed for the evaluation of question wording and provide further recommendations for developing automated procedures to support the integration of a language resources approach into the questionnaire development process. Furthermore, we indicate the study limitations and discuss directions of future research. Lastly, we highlight the originality of this thesis from both a theoretical and practical standpoint.

2 Theoretical background

In this section, we first present the two linguistic fields, corpus linguistics and lexical semantics, focusing on the corresponding resources that are applied in this dissertation, text corpora (Section 2.1) and semantic lexicons (Section 2.2). Second, we present how QUAID used linguistic resources to detect complex survey questions (Section 2.3). Third, we present traditional question evaluation methods based on personal judgment, such as expert reviews and cognitive interviews (Section 2.4). Finally, we present the split-ballot technique as a quantitative survey evaluation method used in the fielding phase of the survey. We also present some previous studies that have used this technique to evaluate question wording variations (Section 2.5).

2.1 Corpus linguistics

Corpus linguistics is the study of language as expressed in large samples of language in natural contexts. Corpus studies are an empirical confirmation of language patterns as normally used by native speakers, and corpora can be used to make 'the decisions that native speakers make subconsciously' (Thomas 2016, 6). Corpora are databases of authentic texts that are compiled for specific purposes, according to specific criteria and aims. The texts used can be anything, from books and newspapers to movie scripts and Internet resources. One of the first corpora to be compiled in English was the Brown University Standard Corpus of Present-Day American English (Kučera and Francis 1961), which was followed by several others in English and other languages.

Every corpus has its advantages and disadvantages. Restricting the analysis to only one would be a limitation to our understanding. Thus, we used three different corpora for the English language:

1. The **British National Corpus (BNC)** (http://www.natcorp.ox.ac.uk) has long been the gold standard for British English; it is a 100-million-word text corpus of written and spoken present-day British English taken from a wide range or sources (Burnard 1995; Leech et al. 2001). It covers the period from 1960 to 1994, although over 93% of the texts are from 1985–1994. It might be slightly outdated, but it has the widest range of sub-genres and includes spoken texts,

which also give us coverage of informal conversations. About 90% of the corpus is written texts, such as excerpts from newspapers, specialist periodicals and journals, books (academic and fiction), letters and memoranda, and essays, while the remaining 10% consists of transcripts of spoken texts. The corpus also contains meta-data on structural properties of the texts.

2. Next, the **Corpus of Contemporary American English (COCA)** (http://corpus.byu.edu/coca/) contains more than 450 million words, so it is about four times larger than the BNC. In fact, it is the biggest freely available genre-balanced corpus of any language (Davies 2010). It covers the period from 1990 to 2012 and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts in (American) English. The BNC and COCA complement each other: the COCA is larger and more up to date, while the BNC has a much wider range of sub-genres and better coverage of informal, everyday conversations.
3. A corpus that is even larger than the BNC and COCA is **enTenTen**, which is created by systematically browsing web content (i.e., web crawling) and is not based on any design specifications; however, it is consolidated by cleaning and de-duplication (Jakubiček et al. 2013). It is part of the TenTen multilingual corpora (https://www.sketchengine.co.uk/documentation/wiki/Corpora/TenTen) that contains more than 10 billion words in different languages (Jakubiček et al. 2013). Although the TenTen English language sub-corpus enTenTen covers many more texts than the BNC and COCA, it has the disadvantage of not being genre-balanced.

In addition, since we used some questionnaires in the Slovenian language, we also used the Slovenian corpus **Kres** (http://www.slovenscina.eu/korpusi/kres), which is a balanced subsample of almost 100 million words from Gigafida: a corpus of written Slovenian that contains more than 1.2 billion words, 77% of which are from newspapers and magazines, while only 6% are from books. Kres is weighted so that 20% are Internet texts, 17% are fiction, 18% are non-fiction, 20% are newspapers, 20% are magazines, and the remaining 5% can be categorised as 'other' (Logar Berginc and Krek 2012).

The standard feature of all listed corpora is the **concordancer,** a search engine that looks through the corpus and lists every single example of the word entered in sortable concordance lines (Thomas 2016). The concordance for a certain query, either a single word or phrase (i.e., string of words), also displays the frequency of that word or phrase. For instance, the frequency of the word 'main' in the BNC is 25,857, while for the word 'principal' it is 5,139. The size of the frequency is relative to the size of the corpora; thus, frequencies of different corpora can be compared only after normalisation.

Concordances and word frequencies can be retrieved from the respective corpora websites, at least for single frequencies. If we are interested in more than that, corpora management software such as **Sketch Engine** (https://www.sketchengine.co.uk/) needs to be used. Sketch Engine is an online corpus software interface that offers more than 200 corpora in 82 languages, including the four abovementioned corpora. It can be used to retrieve different estimates from the listed corpora, mainly concordances but also word sketches.

Word sketches, a distinctive feature of Sketch Engine, are automatic summaries of a word's grammatical and collocation behaviour (Kilgariff 2004). Collocations are sequences of words or terms that co-occur more often than would be expected by chance. Sketch Engine extends the general collocation concept used in corpus linguistics in that collocations are grouped according to particular grammatical relations. In addition, it is also possible to generate a thesaurus and sketch differences which specify similarities and differences between near-synonyms (Kilgariff 2004). For instance, by sketching the difference between the adjectives 'main' and 'principal', we find that the first is more often collocated with 'sewer', 'battery' and 'gas', while the second more often collocates with 'member', 'head' and several other words (see the screenshot in Figure B. 1 in the Appendix).

As we are mainly interested in word frequencies, it should be noted that the word frequency of a certain word is inversely proportional to its rank in a frequency table, a statistical phenomenon known as **Zipf's Law** (Johns 1991). In practice, this means that the most frequent word occurs about twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, and so forth.

Moreover, there is also a similar law for the distribution of meanings over the words in a lexicon, known as **Krylov's Law of Polysemy**. The law states that an increase in the number of meanings conveyed by a word is linked to an increase in its frequency (Krylov 1982). In other words, more frequent words also have more meanings. This might be problematic since the number of meanings is associated with ambiguity, which is one of the determinants of question complexity. This can also be examined using linguistic resources such as semantic lexica, which we cover in Section 2.2.

2.2 Semantic lexicons

Semantics is the study of meanings in language, at the level of words, phrases, sentences and larger units. Apart from the examination of meanings, semantics also studies the relationships between different linguistic units and their compounds, such as synonymy, hypernymy and hyponymy, if we limit ourselves only to those units relevant in the context of this dissertation.

By definition, **synonyms** are equivalent words or phrases that mean exactly or nearly the same thing as another word or phrase in the same language. Synonymy is relative to context, as two words that are synonymous can usually only be interchanged in that specific context. For instance, 'main' and 'principal' share one meaning – 'most important element' – and can be interchanged in contexts such as 'the principal/main river of'. **Hypernyms** are superordinate words that are more generic than a given word, while **hyponyms** are subordinate words that are more specific than a given word. For instance, 'raptor' is one of the hypernyms of 'owl', while 'owlet' is its hyponym.

Semantic lexicons are dictionaries of words labelled with semantic classes so that associations can be drawn between different words. In this dissertation, we use wordnets, which are a standard resource in computational linguistics. The **WordNet** project (http://wordnet.princeton.edu/) is an attempt to organise lexical information in terms of word meanings rather than word forms, as is the practice in conventional dictionaries (Miller 1995). A meaning of a word in WordNet is a *sense* and each sense of a word is in a different *synset* (i.e., a synonyms set). Word meanings are represented by word definitions, and WordNet maps between the many forms and meanings (senses) of words. Some forms have several senses (polysemy), and some senses can be

expressed by several different forms (synonymy) (Miller et al. 1993). True synonyms are rare, so a weaker definition is applied in WordNet – words that denote the same concept and are interchangeable in many contexts. Apart from the English WordNet developed by researchers at Princeton University (Miller 1995; Fellbaum 1998), there are corresponding wordnets in other languages; for example, in Slovenian there is **sloWNet** (Fišer 2009).

2.3 Text corpora approaches in the questionnaire development process

One of the first studies to apply linguistic resources to the process of developing survey questions was conducted by Graesser et al. (2000). They used word frequencies as an indicator of unfamiliarity, which is one of the five classes of problems with survey questions that can be detected by QUAID, a tool they developed as a questionnaire evaluation aid. Specifically, they used the MRC psycholinguistic database (Coltheart 1981) and set the threshold at the familiarity value of 500 or less, and thus left it to the questionnaire designer to decide whether the wording would be problematic for the target population.

In addition, they also used the WordNet lexical database, but only to flag the wording that did not appear in it (Graesser et al. 2000). Aside from that, WordNet is underutilised in the QUAID tool, which is unfortunate, as it is a rich source of information of relations between different word meanings and their forms.

As already mentioned in Section 1.1., another shortcoming is that many of the wordings that are indicated as problematic are actually not problematic – that is, they are false positives (Graesser et al. 2000).

2.4 Standard question evaluation procedures

One of the key questions of this dissertation is how text corpora approaches to questionnaire evaluation relate to standard question evaluation procedures. It should be noted that estimates based on corpora might have different interpretations and various specific problems, so various deficiencies can appear. Therefore, we should not

uncritically follow text corpora and estimate wording (un)familiarity solely on the basis of corpora frequencies.

In fact, traditionally, comprehensibility problems have already been dealt with in several evaluation procedures, particularly with qualitative methods such as cognitive interviews and expert reviews. Other approaches also exist; however, these are less relevant for the specific task of detecting comprehensibility problems. In the following subsection, we describe not only these two approaches but also previous studies that have used them to evaluate linguistic properties of survey questions.

2.4.1 Cognitive interviews

Cognitive interviewing is a field research method that collects data on respondents' cognitive processes while answering survey questions. Its main advantage is that it enables a very intense focus and allows researchers to collect information on survey responses as well as to identify problems that would not otherwise be directly observable (Willis 1999; Snijkers 2002; Mohorko and Hlebec 2013). One of the outcomes of cognitive interviews is that they can clearly point out various comprehensibility issues (Willis et al. 1999). Traditionally, cognitive interviews have been conducted face-to-face, but recently the web mode of conducting cognitive interviews has also been evaluated (Mohorko 2015).

Various techniques can be used, one of the most popular of which is the think-aloud protocol (TAP), where respondents are asked to report whatever comes into their mind as they complete the task of responding to a survey question. However, it is difficult to use this technique in online cognitive interviews and is also less suitable if we are interested in how a respondent comprehends a specific word in a question. In this situation, two other techniques are more suitable: paraphrasing, which asks respondents to repeat the question in their own words; and definitions, which asks respondents to define a certain word in a question. We present these two techniques in more detail in Chapter 5.

To evaluate the effect of changes based on recommendations from cognitive interviewing, Willis (2005) performed a linguistic analysis on a set of questions about drug use. One of the observed characteristics was 'big words', which is another term that can be used to describe unfamiliar words – that is, words with low corpora

frequencies. Using cognitive interviews, Willis was able to improve 10 terms in the study, reducing the number of 'big words' from 53 in the original version to 43 in the improved version. However, several unfamiliar terms remained undetected. In addition, a very serious drawback of the cognitive interview method is that it involves a lot of resources and a lot of time, neither of which many researchers can afford.

2.4.2 Expert evaluation

Another popular method for questionnaire evaluation is expert reviews, which can also be used to detect unfamiliar wordings. Expert evaluation methods utilise the knowledge of professionals in the evaluation of survey questionnaires and can be very efficient, especially in the early development of questions (Lessler and Forsyth 1996; Akkerboom and Dehue 1997).

In the context of question unfamiliarity, Holbrook et al. (2007) used expert reviews to identify problematic linguistic structures in survey questions. Expert reviews have been shown to positively predict problematic questions, but they lack reliability (Willis 1999; Olson 2010; Cerar et al. 2011; Saris 2012; Yan et al. 2012). Moreover, Olson (2010) found variation in identified problems even for experts with similar methodological backgrounds and training, but joined ratings of experts are a good predictor of item non-response and inaccuracies.

Similar to cognitive interviews, expert evaluations involve resources that not everyone who designs a questionnaire is ready to invest. As a result of omitting expensive and time-consuming formal evaluation procedures, many questionnaires remain problematic from a language comprehensibility perspective. On the other hand, word frequency estimates from text corpora are usually inexpensive and, with modern computational linguistic tools, easy and fast to use, at least for the English language. Within this framework, the text corpora approach could be established as an additional pre-testing method in question evaluation, as demonstrated by Graesser et al. (2000, 2006) and Lenzner (2011, 2012).

Nevertheless, a challenge to address here is the relationship between text corpora and expert evaluation. In part, this was already done in two studies conducted by Graesser and his team, who compared the output of QUAID with judgments of three experts in language, discourse or cognition (Graesser et al. 2000) and 12 experts in survey

methods (Graesser et al. 2006). In both studies, they found that the tool is able to identify problems in survey question that experts also identify; however, it also flagged words that were not evaluated as problematic by experts, in particular for unfamiliar terms. If we treat expert reviews as the golden standard, then these words are false positives. Another similar comparison was done by Olson (2010), who compared evaluations of six experts in survey methods to QUAID results, expecting a strong correlation between QUAID and expert ratings for comprehension problems; however, the questions identified as problematic by experts were not similarly identified by the computer tool.

2.5 Field testing of question wordings

Besides pre-fielding methods, such as cognitive interviews and expert reviews (presented in Section 2.4) based on personal judgment, and model-based methods, such as QUAID (presented in Section 2.3) and SQP (presented in the Introduction in Section 1.1.), other question evaluation methods also exist that require survey data collection.

In the following subsections, we first present the indicators of response quality that are usually measured in such studies (Section 2.5.1). Then, we present the split-ballot technique and how it has been used in various studies to evaluate the effect of wording changes.

2.5.1 Response quality indicators

Comprehensibility problems in survey questions can affect survey data quality in various ways. First of all, even if the respondent actually understands the question, the difficulty of comprehending it increases the **response burden** (Bradburn 1978), which also increases **response times** (Lenzner et al. 2010). This can be detected by analysing response latencies and other paradata (Heerwegh 2003; Couper and Kreuter 2013). However, there are several explanations for longer response times, and they do not necessarily mean a greater cognitive effort is being made or a lower response quality will result (Olson and Smith 2015). For instance, the respondent might spend more time reading instructions and/or because of increased efforts to be careful (Fazio 1990). Nevertheless, although the relationship between speed and accuracy is certainly not

linear, ‘the faster individuals respond, the more likely it is that they will make an error’ (Fazio 1990, 80).

Second, the respondent can also react by not responding to a question at all (**item non-response**) or even **dropping out** – in both cases, this creates an error of non-observation. Research shows that survey drop-out, also called survey breakoff, is related to both respondent characteristics such as education and to questionnaire design characteristics (Peytchev 2009).

Third, the respondent might become frustrated and de-motivated by the complexity of the task and start **satisficing** (Krosnick 1996), which means choosing a satisfactory answer rather than making a cognitive effort to respond optimally (Krosnick 1991). As demonstrated by Lenzner (2012), the typical manifestations of satisficing in the case of comprehensibility problems are selecting ‘don’t know’ answers, non-differentiation, and acquiescence, all of which contribute to measurement error.

Fourth, respondents can misunderstand the question and give an incorrect answer (measurement error) (Lozar Manfreda et al. 2002). This can be observed by analysing the **reliability** and **validity** of the instrument.

For all of these issues, many quantitative approaches exist to measure and analyse corresponding problems, but this often requires a large pilot study or even a full study. Finally, we may also use various subjective measures of respondent burden (Hedlin et al. 2005); for instance, at the end of the survey, we can ask respondents if they understood the questions and ask them to evaluate the experience of filling out the questionnaire.

In addition, it should be considered that apart from questionnaire characteristics, respondent characteristics can also affect response quality. Apart from the abovementioned education as a proxy for respondent cognitive skills, other characteristics are also important. For instance, research on Taiwanese English learners has shown that female students are better at learning vocabulary than their male counterparts (Lin 2011). However, a similar and more recent study on Chinese students did not reveal any gender effects (Wei 2014).

2.5.2 Split-ballot question wording experiments

The measures presented in Subsection 2.5.1 can be compared across different versions of the same questionnaire using the split-ballot technique, where the sample is randomly split into two or more equivalent sub-samples. Each part is administered with a different version of the questionnaire with the aim to test a certain causal hypothesis through experimental manipulation (Schuman and Presser 1981).

The method has often been used to evaluate different question wordings. We first present three earlier and prominent experiments of this kind:

1. Duncan and Schuman (1980) presented the results of an experiment with religious beliefs and attitudes with different wordings and contexts. The study was motivated by a visible change induced by altering the wording and context of the question on interest in religion in the Detroit Area Study (DAS). In 1958, the question was worded *'All things considered, do you think you are more interested, about as interested, or less interested in religion than you were 10 or 15 years ago?'.* In 1959, the question read *'All things considered, has your interest in religion grown, remained the same, or decreased over the last 10 or 15 years?'.* The results showed that in controlled experimental conditions, there was no significant difference between Wording 1 and 2. Thus, the differences in the years 1958 and 1959 are supposedly due mostly to context.
2. Rasinski (1989) analysed question wording experiments in the General Social Survey in 1984, 1985 and 1986, and found that even minor changes can affect responses. He focused on the question that asks to evaluate various government spending policies, which reads *'Are we spending too much, too little, or about the right amount on … [Space exploration program; Improving and protecting the environment; Improving and protecting the nation's health; Solving the problems of big cities; Halting the rising crime rate; Dealing with drug addiction; Improving the nation's education system; Improving the conditions of black's; The military, armaments and defense; Foreign aid; Welfare; Highways and bridges; Social Security; Mass transportation; Parks and recreation]?'.* Split-ballot wording experiments were performed in all three years, with three versions in 1984 and two versions in 1985 and 1986. In all experiments for all issues, one type of manipulation consisted of either adding a positive verb before

the label describing the program (e.g., *Social Security* vs. *Protecting Social Security*) or changing the existing verb to a more positive verb or verb phrase (e.g., *Assistance to Big Cities* vs. *Solving Problems of Big Cities*). The wording effect was found in five of these items, but for only two was it significant across years (*Assistance to Big Cities* and *Assistance to Blacks*). In addition, another type of alteration implemented was using a different issue label (e.g., *Halting Rising Crime Rate* vs. *Law Enforcement*). Only three of the five wording manipulations of this kind have shown significant effects (*Law Enforcement, Drug Rehabilitation* and *Assistance to the Poor*).

3. Smith (1987) analysed the same General Social Survey question later analysed by Rasinski (1989, see above), but focused only on one issue: Welfare. The question reads *'Are we spending too much, too little, or about the right amount on ...'* followed by a list of issues that included *'Welfare'* in the first version, which was changed to *'Assistance to the poor'* in the second version and *'Caring for the poor'* in the third. The split-ballot experiment was carried out in 1984 and 1985 (but in the latter year, only in Versions 1 and 2). The percentage of people who responded 'too little' was 24.6% in 1984 and 19.3% in 1985. By altering the issue to *'Assistance to the poor',* an increase of more than 40 percentage points occurred – to 69.3% in 1984 and 64.7% in 1985. Similarly, changing the wording to *'Caring for the poor'* (used only in 1984) increased the percentage to 64%. In both years, the difference was significant ($p < 0.001$). Furthermore, Smith compared the *poor* vs. *welfare* wording in other studies as well. In 1968, the Institute of Life Insurance's Monitoring Attitudes of the Public found that 61% of respondents wanted the government to do more for *'helping the poor'*, but only 32% wanted to do more for *'people on welfare'*. In 1972, a Harris survey found that 62% of respondents wanted an increase for *'helping the poor'*, but only 22% wanted the same for *'people on welfare'*. In both surveys, the exact sample size is not known, so it is not possible to calculate if the difference was statistically significant. In 1976, Yankelovich et al. found that 51% of respondents were in favour of more spending for *'help for the poor',* but only 17.5% were in favour of more spending for *'welfare'* ($p < 0.001$). Repeating the experiment in 1982 showed that 59% of respondents wanted more spending for *'help for the poor'* and 25% for *'welfare'* (sample size unknown). Smith also presented some studies (Hopes and Fears by Gallup,

Harris, and the Institute for Social Research) that compared the *'welfare'* wording with the *'unemployed'* and *'food stamps'* wording, but the differences were less consistent and smaller. All in all, it is obvious that the wording *'welfare'* carries more negative connotations and produces more negative and less generous responses than the *'poor'* wording.

However, the abovementioned studies did not take the issue of wording familiarity into account, while one of our key research questions in this dissertation is the relationship between text corpora frequencies and the quality of survey responses. Namely, we implicitly assume that questions, which use words with higher corpora frequencies, will also result in better survey data quality indicators.

We now present a few past studies which addressed this issue. First, although it does not explicitly mention word frequency, a relevant survey experiment in the context of wording familiarity was conducted by Blasius and Friederichs (2009), who varied the phrasing of seven items using low-brow (everyday) or high-brow (elaborated) language. In fact, it can be assumed that high-frequency words are typical for low-brow language, while low-frequency words are usually used in high-brow language. Response distributions differed significantly for three of the seven items. Blasius and Friederichs suggested using low-brow wording, as it resulted in more diverse responses along different socio-demographic and attitudinal variables.

The effect of low-frequency words on response quality was examined in greatest detail by Lenzner (2011). Six other text-related features of question comprehensibility were also covered – namely, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, complex logical structures, low syntactic redundancy and bridging inferences. Lenzner et al. (2010) initially compared response times, drop-out rates and survey satisficing (i.e., very short response times, neutral responses, acquiescence and primacy effects) in a randomised split-ballot trial, where one version had well-formulated questions, while the other contained suboptimal wordings (for four questions, the manipulation was a low-frequency word). They found that response times in the well-formulated version were significantly longer in 12 out of 28 question items but in only two of the four low-frequency wordings. On the other hand, there were no differences in drop-out rates, item non-response (which was also very low) and satisficing. The results of this study were extended in an eye-tracking study (Lenzner et

al. 2011), which showed that questions with suboptimal text features had a longer fixation time, fixation count and question fixation time.

Lenzner (2012) also examined the effect of question comprehensibility on response quality in more detail in another split-ballot experiment incorporating a bigger sample size and controlling also for verbal intelligence and motivation. The same questionnaire was repeated after two weeks to assess the reliability of responses. This study found that less comprehensible questions reduced response quality (i.e., the number of non-substantive responses and the number of neutral responses). However, only four out of 28 text manipulations consisted of replacing a high-frequency word with a low-frequency synonym (Lenzner 2010).

Although several studies have compared alternative question wordings in a split-ballot experiment, Lenzner's studies presented above are the only experiments, at least to our knowledge, which were based on a psycholinguistic text analysis. Thus, further empirical evidence is needed to better study the effect of wording frequencies on response quality. In particular, more research is needed in different languages, as Lenzner's research was only performed for the German language.

3 Pilot study

As elaborated in previous chapters, the research into cognitive aspects of survey response has indicated unfamiliar terms as one of the psycholinguistic determinants of question comprehensibility problems. In this chapter the estimates of wording familiarity based on text corpora for the English and Slovenian languages were used to detect potentially incomprehensible wordings in two web survey questionnaires for international exchange students at the University of Ljubljana, one for incoming (English) and the other for outgoing students (Slovenian). Two versions of the questionnaire were developed for each language, one with low-frequency (complex) and the other with high-frequency (improved) wordings, and compared in a split-ballot experiment. The results show a lower drop-out rate and a decreased subjective perception of difficulty for the improved language versions.

Here we present a procedure that complements and builds on previous attempts to detect unfamiliar wordings in survey items. The procedure is based on resources used in computational linguistics, a field that uses statistics and the computer sciences to model natural language. Linguistic corpora and lexical databases have had many applications in various fields, both within and outside linguistics. In survey methodology, the only known application is the aforementioned QUAID tool (described in Section 2).

In our procedure, frequencies in text corpora are used as estimates of wording familiarity and lexical databases are used to find alternative wordings. Our approach differs from earlier studies in the fact that we operate with actual numbers (frequencies) from different text corpora. Moreover, when listing alternatives, we use lexical databases instead of thesauri. Thus, we can better distinguish between words that are true synonyms and those words that are only similar. Furthermore, an important difference is that previous studies (Blasius and Friederichs 2009; Lenzner 2010; 2012) were done for German; in contrast, we research the word frequency effect for English and Slovenian. However, it should be noted that we use the two languages as two distinct case studies that are very different and should not be directly compared.

Our aim is to improve survey question comprehensibility by using simpler and clearer wordings based on linguistic corpora and semantic lexica. Through a linguistic analysis of two questionnaires (English and Slovenian), we produced low-frequency and high-

frequency versions that were compared in two split-ballot experiments (one for each case study). In contrast to Lenzner (2010; 2012), we only focused on low-frequency words, so that we could better understand the relationship between word frequencies and response quality. Thus, we were able to produce a greater amount of wording changes from the control (complex, low-frequency) and experimental (improved, high-frequency) versions of the questionnaire. In addition, we introduced subjective indicators of response burden as a measure of response quality.

First, we aim to explore how to use text corpora and semantic lexica to evaluate question wordings and detect unfamiliar words in survey questions. Second, we want to evaluate the effect of wording improvements on response quality. Does using words that have a higher frequency in text corpora improve response quality in terms of response times, breakoff rate, item nonresponse, satisficing and various indicators of response burden?

As already noted in Section 1.4, this preliminary study basically serves as a proof-of-concept study for testing the ideas and a pilot study for preparing a sound and rigorous instrument for the main empirical study in Chapters 4-6.

3.1 Methodology

A web questionnaire used to assess European international student exchange programmes, such as Erasmus, was used for the case study. The questionnaire asked students about their knowledge, skills, and the study environment, focusing on a comparison between their host and home universities. Two questionnaires were prepared, one for incoming and the other for outgoing students at the University of Ljubljana, although they were almost the same. The questionnaire for the outgoing students was translated into Slovenian. Complex, low-frequency wordings were intentionally chosen by the translators.

The questionnaire was 11 screens long and there were one to three questions on each page. The main part of the questionnaire consisted of 21 questions, amounting to 79 items when counting the response options. The word count was 785 for the English version (1,040 when also including the list of countries in the dropdown list) and 791 for the Slovenian version (1,046 including the list of countries in the dropdown list). For

both versions, we made a list of different nouns, verbs, adjectives and adverbs that appear in the questionnaire and manually searched for their synonyms and other related words for the corresponding sense in WordNet (for English) and sloWNet (for Slovenian). Some of the new words were excluded because they did not sound natural in the context sentence. In most cases, we were limited only to single words but for three wordings in English *('critical assessment', exam mark', and 'subject field'*) and for four wordings in Slovenian (*'delo s tabelami', 'editiranje tekstov', 'ekstrakurikularne aktivnosti', and 'študijski materiali'*) we considered phrases.

For words that have at least one synonym in wordnet, we manually searched the word frequencies of the original wording and other alternatives in three English corpora (BNC, COCA and enTenTen) and one Slovenian corpus (Kres) (all four described in Section 2.1). Where more than one alternative was possible, we kept only the one with the highest wording frequency. In some cases, we replaced the original wording (in the control version) with a lower frequency word to make it more complex. As mentioned, the Slovenian version was already translated in such a way that it contained a lot of low-frequency wordings.

Following the described procedure, we were able to find an alternative wording with a higher frequency in at least one of the corpora for 23 words in the English version. Table 3.1 shows the wordings we used and compares the wording frequencies in BNC; COCA and enTentTen. The last column (N) shows the number of times a wording change was made in the questionnaire.

Table 3.1: Words used in the control and experimental groups and their frequencies according to the BNC, COCA and enTenTen corpora

Control group	BNC	COCA	enTenTen	Experiment.al group	BNC	COCA	enTenTen	N*
accessibility	317	1,665	125,653	availability	1,912	6,322	329,250	3
acquired	6,354	10,357	360,800	learned	23,394	53,748	917,521	3
adequate	3,571	10,472	362,254	enough	25,635	172,956	3,626,045	8
approximately	2,837	17,082	736,000	about	144,554	1,247,800	26,133,577	4
(critical) assessment	7,602 (21)	26,268 (69)	74,3430 (1,318)	(critical) evaluation	2,983 (37)	15,140 (101)	507,716 (1,635)	1
categorise (categorize)	323 (20)	4 (757)	13,841 (73,886)	classify	1,553	1,293	181,497	1
circumstances	11,009	20,187	645,617	conditions	23,742	45,114	1,627,180	1

Control group	BNC	COCA	enTenTen	Experiment.al group	BNC	COCA	enTenTen	N*
completed	9,711	21,354	899,121	finished	11,977	30,787	899,121	1
constituent	1,512	1,284	98,596	part	65,773	224,094	7,697,844	1
engage	4,258	13,299	372,098	included	34,753	48,249	1,518,267	1
enrol (enroll)	500 (6)	22 (1,901)	11,634 (245,561)	enter	14,141	21,576	1,821,556	1
evaluate	2,238	8,412	593,651	rate	1,418	58,799	4,207,393	9
fellows	3,155	2,486	69,669	colleagues	7,209	22,071	321,680	1
furthermore	2,918	11,600	482,069	moreover	4,327	17,911	440,230	1
impacted	65	1,528	90,358	affected	5,923	17,754	619,361	6
instructors	850	4,671	119,687	teachers	19,744	77,840	779,893	5
laboratory	3,748	13,328	358,032	lab	942	14,870	382,042	3
(exam) mark	6,139 (6)	55,738 (0)	1,787,037 (174)	(exam) grade	2,525 (1)	24,817 (6)	768,584 (387)	1
obligatory	327	916	33,042	mandatory	968	5,289	168,921	1
obtained	12,382	13,999	483,251	received	24,111	49,366	1,731,037	2
oral	1,898	9,332	327,908	spoken	25,788	13,161	234,477	3
prerequisites	88	529	41,757	requirements	9,234	1,5847	1,367,669	3
subject field	62	38	1,384	field of study	81	263	15,064	1

*N = number of appearances in the questionnaire (for full questions see Appendix A)

We changed 23 different wordings in the English version but some appeared just once in the questionnaire, while others appeared several times – the most frequent were *'adequate'* (8) and *'evaluate'* (9). As mentioned, there were three cases where we examined a phrase and not a single word. If we had considered the single-word frequency for *'assessment'* and *'mark'*, we would have arrived at a different decision.

The wording frequencies in the three different English corpora are usually consistent – if a word has a relatively low frequency in one corpus it is also low in the other two. However, there are exceptions. For instance, the words 'evaluate 'and 'laboratory' are less frequent than *'rate*' and *'lab'* according to the COCA and enTenTen but more frequent according to the BNC. Similarly, *'furthermore'* and *'oral'* are more frequent than *'moreover'* and *'spoken'* according to the BNC and the COCA, but less frequent according to enTenTen.

For the Slovenian version we were able to find 39 cases for which there was an alternative wording with a higher frequency than the original version. Table 3.2 presents the frequencies according to the Kres corpus.

Table 3.2: Words used in the control and experimental groups and their frequencies according to the Kres corpus

Control group	Kres	Experimental group	Kres	No. of changes
absorbirati	410	pridobiti	15,585	1
adekvaten	51	ustrezen	18,130	3
aspekt	628	vidik	8,448	1
definitiven	44	odločen	3,965	1
delo s tabelami	4	delo s preglednicami	14	1
editiranje tekstov	0	urejanje besedil	46	1
ekstrakurikularne aktivnosti	0	dodatne dejavnosti	77	1
enormen	136	ogromen	5,016	1
evalviranje	69	vrednotenje	2,510	1
evalvirati	95	oceniti	6,227	6
infrastruktura	3,280	oprema	16,575	2
inozemstvo	49	tujina	10,987	1
institucija	6,832	ustanova	7,495	2
kapaciteta	1,150	sposobnost	12,756	3
klasificirati	126	uvrstiti	5,177	1
komuniciranje	2,079	sporazumevanje	1,172	1
komunikacijski	2,450	sporazumevalen	178	1
konverzacija	98	pogovor	19,674	1
kriterij	4,868	merilo	7,283	1
kurz	101	predmet	24,579	2
kvaliteta	3,443	kakovost	12,034	2
kvantiteta	141	obseg	8,344	3
lokacija	5,548	kraj	28,826	1
lokalec	60	domačin	4,868	1
neadekvaten	10	neustrezen	1,555	1
nivo	4,004	stopnja	21,403	2
oralen	427	usten	3,231	2
participacija	457	sodelovanje	24,449	3
pedagog	1,054	učitelj	15,010	4
prezentacija	199	predstavitev	9,676	1
prezentirati	36	predstavljanje	593	1
razpoložljivost	238	dostopnost	1,136	3
rigoroznost	2	strogost	319	1
socialen	24,619	družaben	3,124	1
sumaren	15	v celota	8,656	1
(študijski) material	14,838 (8)	(študijsko) gradivo	9,014 (101)	1
timski	635	skupinski	3,054	1
verziran	2	izučen	134	1
verziranost	1	spretnost	3,531	2

In the Slovenian version, 39 different wording changes were made. Most of them appeared only once but some appeared several times in the questionnaire, most

frequently *'evalvirati'* (6) and *'pedagog'* (4). As mentioned, there are four wordings where we looked up the phrase and not a single word. The word frequency for *'material'* is lower than *'gradivo'* so the decision would be different if we had focused on individual wordings. That might also be the case for some other words in the table, but this is a point for further exploration.

There are three words for which the wording in the control version actually has a lower frequency than in the experimental version (*'komuniciranje', 'komunikacijski'* and *'socialen'*). We decided to allow this exception for stylistic reasons: many of the words in the control group are words of foreign origin and their alternatives are Slovenian synonyms of these words. Thus, the control version has a style that employs a lot of foreign words, while the experimental version uses domestic alternatives. These three words are also of foreign origin and we thus decided to have them in the complex control version and use the more Slovenian wordings in the experimental version.

At the end of both the control and experimental versions of both questionnaires (English and Slovenian), we included a block of questions that measure respondent satisfaction and questionnaire difficulty. The following questions were asked:

- *How much did you enjoy completing the questionnaire? A great deal, A lot, A moderate amount, A little, Not at all.*
- *How difficult was it for you to interpret the meanings of questions in this questionnaire? Extremely difficult, Very difficult, Moderately difficult, Slightly difficult, Not difficult at all.*
- *How difficult was it for you to generate answers to the questions in this questionnaire? Extremely difficult, Very difficult, Moderately difficult, Slightly difficult, Not difficult at all.*
- *How many times did you not understand a certain word in a question? Please give at least an approximate answer. If there were no such words, please write 0.*

In addition, we were interested in the respondents' multitasking behaviour and assumed that respondents are less prone to perform other activities (e.g., visiting other websites) if the questionnaire is less demanding for them. We measured multitasking with two questions, one for multitasking on electronic devices, and the other for other

multitasking activities. Both questions had eight different activities listed and multiple answers were possible (a check-all-that-apply format). The question wording was: *What, if anything else, have you been doing on any electronic device while responding to this survey?* And: *What, if anything else, have you been doing while responding to this survey?*

3.2 Results

The study was carried out in April and May 2014 with Erasmus exchange students at the University of Ljubljana. The survey invitation (and one reminder) was sent to 1,147 incoming (international) and 917 outgoing (Slovenian) students. Following a random allocation, about half the respondents were allocated to the control (complex) and half to the experimental (improved) version that we described in the previous section. In total, 230 (20%) incoming students and 205 (22%) outgoing students started responding to the survey. The incoming students were responding to the English version and the outgoing students to the Slovenian version.

The incoming students who responded to the English version came from 27 different countries, mostly European. The largest group were Spanish students (11% of the respondents). No students were from an English-speaking country but five reported they are native speakers of English. It should be noted that for most of the respondents English was not their first language, which makes them more prone to comprehension difficulties.

We observed differences in five indicators of response quality: item non-response, drop-outs, straightlining, response time (average and median), subjective burden, and multitasking. Drop-outs are those who left the survey between the second and penultimate page of the questionnaire. The item non-response rate was computed by counting the number of items (out of 64) that were left blank. Straightlining is a manifestation of satisficing and is defined as always selecting the exact same response in a matrix question, either the middle point or another answer in the matrix. We computed straightlining for the four matrices that had more than three items: Q4 has eleven items and three response options (inadequate, just adequate, more than adequate), Q8 has eight items and five response options (much lower, lower, approximately the same, higher much higher), Q14 has five items and six response

options (no information, a little, a moderate amount, a lot, a great deal of information) and Q15 which has eight items and five response options (no information, a little, a moderate amount, a lot, a great deal of information). Drop-outs were removed when computing the item non-response and straightlining. In addition, when computing the average and median response we removed item non-respondents and outliers (those who took more than one hour to respond).

Subjective burden was measured with four indicators, namely: enjoyment in completing the questionnaire, the difficulty of interpreting the meanings of questions, the difficulty of generating answers to questions, and the amount of times the respondent did not understand a certain word. Even if the variables have ordinal measurement scales, we assumed it is an interval scale and computed averages. Multitasking was measured with four questions, but we only analyse the first two: multitasking on a computer (or other device) and multitasking without a device. For both, we counted the number of boxes the respondent ticked but classified them as an on- or off-computer multitasker where they ticked at least one.

3.3 Analysis

We applied different statistical tests for different measures. A chi-square test was conducted for drop-outs and straightliners that were measured as a dummy and the percentage of those who were classified as a drop-out or straightliner is shown. For all other measures we calculated averages and medians. For averages we carried out Student t-tests for independent samples, while for medians we used the nonparametric Mann-Whitney U test. We also computed Cohen's d and r as effect sizes for all tests (Cohen 1988). The results for the English version are presented in Table 3.3 and for the Slovenian version in Table 3.4.

Table 3.3: Comparison of the control and experimental English versions

Indicator		Complex	Improved	Test	Effect size	n1	n2
Drop-out rate (%)		30.8	20.0	Chi^2=3.53* (p=0.06) df=1	d=0.25 r=0.12	120	110
Item non-response	Average	4.3	4.4	t=-0.5 (p=0.88)	d=-0.08 r=0.01	83	88

Indicator		Complex	Improved	Test	Effect size	n1	n2
	Median	4	4	U=3357 z=0.29 (p=0.39)	d=0.04 r=0.02	83	88
Straightlining[1] (%) (always selected the exact same response in a matrix)		16.9	21.6	Chi^2=0.61 (p=0.43) df=1	d=0.10 r=0.05	83	88
Response time[2] (sec)	Average	834.1	807.7	t=1.07 (p=0.29)	d=0.16 r=0.08	80	84
	Median	636.5	721.0	U=3031.5 z=1.17 (p=0.28)	d=0.18 r=0.09	80	84
Enjoyed responding survey (1 – Extremely 5 – Not at all)	Average	3.1	3.1	t=0.32 (p=0.75)	d=0.05 r=0.03	83	88
	Median	3	3	U=3437.5 z=-0.66 (p=0.25)	d=0.10 r=0.05	83	88
Difficulty of understanding questions in this survey (1 – Extremely difficult, 5 – Not difficult at all)	Average	4.6	4.7	t=-0.42 (p=0.68)	d=0.06 r=0.03	83	87
	Median	5	5	U=3521 z=0.22 (p=0.39)	d=0.03 r=0.02	83	87
Difficulty of providing answers in this survey (1 – Extremely difficult, 5 – Not difficult at all)	Average	4.5	4.5	t=-0.34 (p=0.73)	d=0.05 r=0.03	82	88
	Median	5	5	U=3595 z=-0.04 (p=0.48)	d=0.01 r=0.00	83	87
How many times certain words were not understood (numeric input)	Average	0.5	0.8	t=-1.40 (p=0.16)	d=0.23 r=0.11	76	78
	Median	0	0	U=2812.5 z=0.55 (p=0.29)	d=0.09 r=0.04	76	78
Multitasking **on** computer/device (% who indicated at least one activity)		38.3	37.3	Chi^2=0.03 (p=0.89) df=1	d=0.02 r=0.01	120	100
Multitasking **off** computer (% who indicated at least one activity)		16.9	21.6	Chi^2=0.22 (p=0.72) df=1	d=0.06 r=0.03	120	100

[1] Drop-outs removed

[2] Drop-outs and item non-responses removed

The group that responded to the improved English version had a lower drop-out rate (20%) than the complex (control) version by almost 10 percentage points (30.8%). It is an important difference and turns out to be significant at the 0.06 level (Chi-square = 3.53) and although the sample is small, there is a small power effect (Cohen's d = 0.25).

Furthermore, we also checked the drop-out per page. Most of the drop-outs occurred on the first page: 20 (17%) in the low-frequency version and 12 (11%) in the high-frequency version. Note that there was one wording change on this page ('constituent' vs 'part'). The remaining drop-out happened on the second page or later: 17 cases (14%) in the low-frequency and 10 cases (9%) in the high-frequency version. Per page differences go in the direction of our hypothesis; however, the cell sizes are too small to generalize.

On the other hand, there were no differences in item non-response, straightlining, response times, and in the subjective burden indicators. Either the sample size was too small or changing the 23 wordings does not have any effect on different measures of response quality (other than drop-out). In contrast, in the Slovenian version, where 39 wordings were changed, the results are somehow different (Table 3.4).

Table 3.4: Comparison of the control and experimental Slovenian versions

Indicator		Complex	Improved	Test	Effect size	n1	n2
Drop-out rate (%)		34.0	28.3	Chi^2=0.77 (p=0.38) df=1	d=0.12 r=0.06	106	99
Item non-response	Average	3.5	3.8	t=-0.76 (p=0.45)	d=0.13 r=0.06	70	71
	Median	4	5	U=2285.5 z=0.82 (p=0.41)	d=0.13 r=0.07	70	71
Straightlining[1] (%) (always selected the exact same response in a matrix)		30.0	31.0	Chi^2=0.01 (p=0.90) df=1	d=0.01 r=0.00	70	71
Response time[2] (sec)	Average	668.2	682.7	t=-2.16 (p=0.80)	d=0.04 r=0.02	67	69
	Median	549.0	582.0	U=2309 z=-0.28 (p=0.77)	d=0.05 r=0.02	67	69
Enjoyed responding survey (1 – Extremely, 5 – Not at all)	Average	3.2	3.4	t=-1.35 (p=0.17)	d=0.23 r=0.11	70	71
	Median	3	3	U=2222 z=1.08 (p=0.28)	d=0.16 r=0.08	70	71
Difficulty of understanding questions in this survey (1 – Extremely difficult, 5 – Not difficult at all)	Average	4.0	4.8	t=-6.17*** (p=0.00)	d=0.60 r=0.32	70	71
	Median	4	5	U=1278*** z=4.98 (p=0.00)	d=0.84 r=0.39	70	71

Indicator		Complex	Improved	Test	Effect size	n1	n2
Difficulty of providing answers in this survey (1 – Extremely difficult, 5 – Not difficult at all)	Average	4.4	4.6	t=-1.76* (p=0.08)	d=0.30 r=0.15	70	71
	Median	4	5	U=1942** z=2.24 (p=0.03)	d=0.35 r=0.17	70	71
How many times certain words were not understood (numeric input)	Average	1.3	0.1	t=5.36*** (p=0.00)	d=0.57 r=0.27	70	71
	Median	1	0	U=2178 z=-0.57 (p=0.57)	d=0.09 r=0.04	70	71
Multitasking **on** computer/device (% who indicated at least one activity)		27.4	30.3	Chi^2=0.22 (p=0.65) df=1	d=0.07 r=0.03	106	99
Multitasking **off** computer (% who indicated at least one activity)		8.5	6.1	Chi^2=0.45 (p=0.60) df=1	d=0.09 r=0.05	106	99

[1] Drop-outs removed
[2] Drop-outs and item non-responses removed

Improving the wording of the Slovenian questionnaire decreased the impression of difficulty of understanding questions from 4.0 to 4.8 points (t=-6.17, p=0.00) and impression of difficulty of providing answers from 4.4 to 4.6 points (t=1.76, p=0.08). The differences are confirmed also by the Mann-Whitney U test: for both the difficulty of understanding the question (z=4.98, p=0.00) and difficulty of providing an answer (z=2.24, p=0.03) the median increases from four to five in the improved version (meaning less difficulty). The effect sizes for the difficulty to understand is intermediate (d=0.60) for the t-test and high for the z-test (d=0.84), while for the difficulty of providing an answer the effect is small both for the t-test (d=0.30) and z-test (0.35). Looking only at the average value, there is also a significant difference in the number of words not understood from 1.3 to 0.1 (t=5.36, p=0.00) but it is not confirmed by the nonparametric median test and the sample size is too small to give it statistical power.

On the other hand, the decrease in drop-out rates is smaller than in the English version, less than five percentage points (from 34% to 38.3%) and not even close to significant (Chi-square = 0.77, p = 0.38). Moreover, there are no significant differences in item non-response, straightlining and response times.

Finally, it should be noted that three out of the 64 total changes (of 39 different wordings) in the Slovenian version were not in line with other changes. While most of

the changed wordings in the (supposedly) improved version were words with a higher frequency than in the control version, those three changes went in the opposite direction. However, they appeared towards the end of the questionnaire and present a minimal (4%) change compared to all changes that were done in the proper direction.

3.4 Discussion

In this section, we first briefly recall the challenges of using text corpora and lexical databases to improve survey question wording, which is an under-researched topic in the field of questionnaire design. In particular, we summarized Lenzner's (2010; 2012) research on the effect of different text features on response quality and we outlined an empirical study based one of his research. However, we only concentrated on the effect of wording frequencies, allowing us more focus, and instead of German we applied our study on two other languages: English and Slovenian.

The study confirmed that the specific action of improving question wording by using words with higher frequencies can have a certain effect on some indicators of response quality. Although the results are somewhat different from those of Lenzner, we also confirmed some basic tendencies from his studies. Let us summarize the key findings.

First, as in Lenzner's studies (2010; 2012) we were not able to observe any difference in item nonresponse and satisficing, neither in the English, nor in the Slovenian questionnaire. However, it should be noted that Lenzner looked into four indicators of satisficing (very short response times, neutral responses, acquiescence, and primacy effects), while we looked only into one (straightlining).

Second, although Lenzner hypothesized that word frequency might have an effect on drop-out rates, his evidence showed no significant differences for this indicator. In our experiment, on the other hand, we observed a small effect on drop-out rates, which was confirmed also by the power analysis. The replacement of 23 wordings in the English version with alternative wordings of higher frequency reduced the drop-out rate by almost 10 percentage points. Moreover, we can see a lower drop-out also for the Slovenian language version, as the 39 wording changes decreased the drop-out rate by almost five percentage points; however, the Slovenian findings are not significant and

cannot be generalized due to the small sample size. Nevertheless, the same tendency as in the English version was confirmed.

Third, we were not able to observe significant changes in response times, which is one of the main results of Lenzner's research (2010; 2012). Although we actually observed a small difference in response times between the control and improved versions of the questionnaires for both languages, the differences were small and not significant. In any case, the sample size is too small to give these results any statistical power.

Fourth, what is novel in our experiment is that we also looked into some subjective measures of response burden, namely how much the participants enjoyed responding, the difficulty of understanding the questions, the difficulty of providing answers, and the number of times a certain word was not understood. For the English language questionnaire there were no effects, but for the Slovenian language questionnaire we observed a moderate effect for the difficulty of understanding and a small effect for the difficulty of providing answers. Also, there was a significant difference for the average number of times the respondents did not understand a certain word; however, the power analysis did not confirm the effect for the latter.

The differences in some research findings between our study and Lenzner's study can be primarily explained by differences in the methodological approach, i.e., per item observations in Lenzner's study vs. observing the aggregated effect of a series of changes. In addition, the specifics in the study populations, the language, and the questionnaire are just as important in explaining the differences.

With respect to certain differences in the strength of conclusions between our English and Slovenian study, it should be emphasized that they differed in the nature and amount of wording alternations. Moreover, as stated in the introduction, the two experiments are distinct case studies that represent two different populations and two different languages, which should not be directly compared. While the perceived lower number of incomprehensible words and the decreased perception of difficulty in the improved Slovenian questionnaire could be explained by the higher amount of wording improvements in the Slovenian version, there is no immediate explanation, except for some cultural effects, for the less pronounced decreased in the drop-out rate, compared to the English version.

3.5 Conclusions and study limitations

The pilot study confirmed the basic findings of Lenzner (2011) that word frequencies can have some effects on question comprehension and response quality. However, due to the differences in methodological approaches, some effects found in our study were different. While Lenzner found an impact on response times, we found effects on drop-out rates and on the subjective perception of response burden. Nevertheless, both studies found no effect on item nonresponse and satisficing.

However, due to design limitations and small sample size in this study, it is difficult to accurately evaluate the specific effect of question wording on different response quality indicators.

In any case, this preliminary pilot study confirmed the conceptual relation between wording frequencies and survey data quality indicators. It also outlined the direction for the work in proceeding chapters. It showed the importance of careful selection of questions items and cases word wordings, which are suitable for variations. In addition, it confirmed that the language resources used are suitable for this purpose, as well as the fact that strings of the words and the related context need to be observed, instead of only single word frequencies.

Of course, the current study has some conceptual and methodological limitations:

- The first limitation of the study is the relatively small sample size, which comes from only one university. Since we want to estimate small proportions (i.e., the percentage of drop-outs), a sample of at least 400 units per group is needed. Implementing the study on a larger population (other university, general population) would also empower the results. However, the relatively narrow population and small sample size does not jeopardize the internal validity of the findings, which certainly exposes a potential for the high effect of word frequency on response quality.
- Second, the design of the questionnaire does not allow for a very accurate measurement of item response times. Since there is more than one item on each page, it is impossible to measure the time needed to respond to a specific item. A paging design shall be used in future experiments to enable the calculation of question response times.

- Third, the number of experimental groups is another limitation. Two groups might be enough to evaluate all differences only as one integrated factor, but is not sufficient to study the effect of more specific factors, such as the nature and origin of alternative words, the specific effects of single words, the effect of the total number of words changed, the extent of change (moderate vs. high difference), the topic of the questionnaire, and also the role of specific factors of the target population (culture, language, socio-demographics). There are almost countless variations and requests for additional experimental cells.

4 Comparing expert evaluations and text corpora

In this chapter, we explore two approaches for evaluating the comprehensibility of questions. Relying on computational linguistics research, we first use wording frequencies based on large text corpora as the estimates of wording familiarity. Supposedly, wordings with a higher frequency are more comprehensible to the respondent. Next, we evaluate whether these wording recommendations are compliant with expert reviews, as one of the common question evaluation methods.

Two very diverse sets of questions are used as case studies for these comparisons: a selection of nine (questions) items on wages and working conditions, and a selection of 12 (questions) items on attitudes towards terrorism. For each item, we identify the selected word (or sequence of words), typically one word in each item, except for one item where two words are studied. For the selected word, we also define a set of alternative words (or sequence of words). We thus have 20 cases where the original and alternative wordings are the subject of comparisons, so that both sets of selected words and corresponding alternatives are evaluated with a text corpora approach and also with expert reviews (about 50 for each study).

As presented in Section 2, expert reviews have already been used to detect unfamiliar wordings and other problematic linguistic properties of survey questions. However, previous studies found some mismatches between QUAID, the automated tool for question evaluation, and expert reviews (Graesser et al. 2000; Graesser et al. 2006; Olson 2010). Moreover, results varied across different experts, even in cases of similar background characteristics (Olson 2010).

Instead of using the QUAID tool, in the present study we compare expert evaluations directly to text corpora frequency estimates in larger corpora, without being limited to a certain threshold in just one corpus. Moreover, instead of comparing how successful the two approaches are in detecting problems, we focus on how experts evaluate a series of alternative wordings of the same question that differ in their corpora frequency.

In addition, we also address the effect of linguistic skills: Are evaluations by non-native speakers the same as those of native speakers, and if not, whose evaluations better correspond to results of the text corpora approach? This is important because in practice

it is often the case that the questionnaire designer is not a native speaker. Presumably, there is an effect, at least for so-called 'false friends' – that is, 'pairs of words that are the same or almost the same in two languages but whose meanings and/or usage differ' (Thomas 2016, 177). For instance, the word 'sympathetic' (see items P4.2 and P4.3 in Section 4.2.2) has a different connotation in German, Italian, Spanish and many other languages. Thus, native speakers of these languages might understand such words differently than intended by the questionnaire designer.

In the next section, we first present the questions we use as case studies. We describe the process of selecting the questions and also the selection of alternative wordings, which are then included in both approaches. We then present the results, first for the text corpora approach and second for expert evaluations. Then, we compare results of native and non-native speakers. Finally, we compare the two approaches and discuss the results.

We may also add that the empirical implementation of the proposed wordings in the corresponding split-ballot survey experiments is not discussed here, but rather in Chapter 6. We focus here only on comparisons of the two approaches in the stage of preparing the questionnaire.

4.1 The selection of items, cases and alternative wordings

In this section, we explain how we selected the question items and corresponding cases of wordings that were then evaluated, first with the text corpora approach and then with expert evaluation. The basic criterion was to select question items that contain unfamiliar wordings which have at least one synonym in WordNet with a different corpora frequency.

However, in making a purely random selection of wordings (to be compared with their alternatives), we would probably not find the most appropriate cases for this study. Thus, a careful selection process was needed in order to devise a list of question items with relevant wordings that would enable comparisons of the two approaches and help discovering potential effects. It should be noted, however, that the selection is not representative of a population of survey questions. The two case studies – where we compare the corresponding performance of text corpora and expert evaluations in

detecting unfamiliar wordings – rather serve here as proof of concept, where we can study whether corpora can be used in improving question wording but also to what extent the corpora approach can replace expert evaluation. For this reason, we needed a set of examples that are typical but also operationally suitable for our research.

We thus took the two sets of questions from two different surveys as case studies to explore how to improve question wording with linguistic resources. The first case study is a selection of eight questions (nine question items) from the WageIndicator survey questionnaire. In this study, we evaluated seven different words, but one of them appeared in three different contexts (in three items). So in total, there were nine different strings of words (i.e., word sequences). The second case study is a selection of eight questions (12 question items) from the Pew Research Center database of polling questions. In this study, we evaluated 12 different words, but one of them was used in two different contexts (in two items). So in total, there were 13 different strings of words. In fact, one of the items (P5) contained two different words that we evaluated.

The words were selected based on a list of wordings with corresponding synonyms and other alternatives retrieved from WordNet, which were then checked against different corpora. There were some differences in the process of selection between the two case studies because, in the Wageindicator case, we evaluated the whole questionnaire and made a selection; while in the PEW case, we were looking at a much larger database that contained questions from several surveys, and we made a selection of question items from different questionnaires on the same topic.

In the first step, when making the selection of the items, we considered basic frequencies of potential words, which were then entered into the study. For this purpose, we used the BNC rather than COCA or enTenTen, because a database of BNC wordings (with a frequency of 800 and higher) is available to download, while the wordings in the other two corpora can only be accessed by several specific queries. However, in the evaluation step, the two other corpora were more useful as they enabled the retrieval of frequencies for strings of words, while the BNC is limited only to single frequencies. In the following subsections, the search for alternative wordings and the evaluation of both questionnaires with the COCA and enTenTen corpus is presented.

4.1.1 The selection process for the Wageindicator Survey

The Wageindicator survey is a continuous, multi-lingual, multi-country, non-probability web survey that is used to collect data on wages and labour conditions in more than 80 different countries (Tijdens and Osse 2014). It started in 2000 in the Netherlands and is run in cooperation by the Wageindicator Foundation, a non-profit Dutch organisation, and the Amsterdam Institute of Advanced Labour Studies (AIAS). A list of variables and values is available in the Wageindicator Codebook (Tijdens and Fabo 2014).

We received the full questionnaire from AIAS and evaluated it with linguistic resources. It has 522 question items and the total word count of the English version of the questionnaire is 6,181 words. However, the structure of the questionnaire is quite complex with several filter questions, so most respondents only respond to a part of the questionnaire.

In the first step, we examined each question by searching for synonymous wordings for the main word (subject) in the item. In case there were no synonyms for that word in WordNet, we looked up synonyms for other words in the item; if there were either no synonyms or only a few synonyms, we looked up other alternative wordings (i.e., similar words, hypernyms, hyponyms). Thus, we generated a list of 203 wordings (nouns, verbs, adjectives and adverbs) for which we found at least one alternative. Some of the words appeared in more than one question item.

In the next step, we further examined these 203 words and their alternatives by looking up their word frequencies in the BNC. In 75 cases, at least one of the alternatives had a higher wording frequency than the original wording. We reviewed them again and decided to exclude 20 wordings because they did not fit into the context, 25 wordings because the difference between the original and the alternative was not very big, and 18 wordings because the original frequency was high enough. Thus, only 12 wordings were relevant for further analysis.

Next, we excluded three words because they appeared in questions with long lists of response options and were not very central in the respective questions. Thus, we considered them less interesting for the expert evaluation and kept only a final selection

of nine items (eight question wordings and one response option) to be evaluated with different methods. The selected wordings are underlined:

1. What kind of employment contract do you have?
2. Is your organisation domestic or foreign-owned?
   - Wholly domestic-owned;
   - Partly domestic owned, partly foreign owned;
   - Wholly foreign-owned.
3. Do you usually work the number of hours laid down in your contract?
4. How often does your job involve solving unforeseen problems on your own?
5. To what extent do you agree with the following statements?
   5.1 I have sufficient energy to do my job.
   5.2 I have sufficient support from my supervisor.
6. My job is sufficiently varied.
7. Machines/equipment are in a good state of repair.
8. Staffing levels are sufficient.

Each of the underlined wordings was looked up in the WordNet online tool and the results are presented in Table 4.1.

Table 4.1: Synonyms and other alternative wordings found in WordNet: Wageindicator questions

Case	Query*	Wordnet Synset	Definition
W1	kind	**kind**, sort, **form, variety** (n)	a category of things distinguished by some common characteristics or quality
	[hyponym]	**type** (n)	a subdivision of a particular kind of thing
…	[hyponym]	… [15 other hyponyms] (n)	… [various definitions]
	[hypernym]	category (n)	a general concept that marks divisions or coordinations in a conceptual scheme
W2	wholly	**wholly, entirely, completely, totally, all,** altogether, whole, right (adv)	to a complete degree or to the full or entire extent
W3	laid down	l**ay down, establish, make** (v)	institute, enact, or establish
	troponym	**set**, mark (v)	establish as the highest level or best performance
	hypernym	make, create (v)	make or cause to become
W4	unforeseen	**unanticipated, unforeseen, unseen, unlooked-for, out of the blue** (adj)	not anticipated
	similar to	**unexpected** (adj)	not expected or anticipated
W5	sufficient	**sufficient** (adj)	of a quantity that can fulfil a need or requirement but without being abundant
	see also	ample (adj)	more than enough in size or scope or

Case	Query*	Wordnet Synset	Definition
			capacity
	similar to	**adequate**, **enough** (adj)	sufficient for the purpose
	similar to	comfortable (adj)	sufficient to provide comfort
W6	sufficiently	**sufficiently** (adv)	to a sufficient degree
	no relation	**adequately** (adv)	in an adequate manner or to an adequate degree
	no relation	**enough**, plenty (adv)	as much as necessary
W7	(state of) repair	**repair** (n)	a formal way of referring to the condition of something
	hypernym	**condition**, status (n)	a state at a particular time

*Note: The column 'Query' represents the word that we entered in the WordNet browser (available at http://wordnetweb.princeton.edu/perl/webwn). An example of query results for case W1 is presented in the Appendix (see Figure B.2). Moreover, it is also possible to click and open additional layers to retrieve also hyponyms, hypernyms and other alternative wordings (see Figure B.3 in the Appendix).

First, we found the corresponding set of synonyms (synset). In most cases, there were relevant synonyms in the synset, i.e., 'kind, sort, form, variety' (W1), 'wholly, entirely, completely, totally, all' (W2), 'lay down, establish, make' (W3), 'unforeseen, unanticipated, unseen, unlooked-for, out of the blue' (W4). Second, we looked for other similar words based on the 'see also' and 'similar too' relations in WordNet. Third, for nouns and verbs, we also looked for hyponyms (troponyms in the case of verbs) and hypernyms. The final selection for examinations was made based on their wording frequencies and evaluation of plausibility: wordings with a definition that did not suit the context or with very low frequencies were not selected.

The words in bold are those that we selected for further examination using wording frequencies from two different corpora and expert reviews (Sections 4.2.1 and 4.2.2). Specifically, we skipped wordings that we did not consider relevant enough in the context. For the noun 'kind' (W1), we took two of its synonyms in the synset (form and variety) and the hyponym 'type' – the only one out of 16 with a suitable definition. For the adverb 'wholly' (W2), we took four of the seven wordings in the synset (entirely, completely, totally and all). For the verb 'laid down' (W3), we took the two synonyms in the synset (establish and make) and the troponym 'set'. For the adjective 'unforeseen' (W4), we took four synonyms in the synset (unanticipated, unseen, unlooked-for, and out of the blue) and one similar wording (unexpected). For the adjective 'sufficient' (W5.1, W5.2 and W8), we took two similar words (adequate and enough). For its adverb 'sufficiently' (W6), there were no related wordings in WordNet; thus, we formed

them based on the adjectives in the previous case: ‘adequately’ and ‘enough’. For the noun ‘repair’ (W7), we took one hypernym (condition).

4.1.2 The selection process for PEW questions

The Pew Research Center (PEW) is a non-profit, non-partisan American research institute that provides information on social issues, public opinion and demographic trends shaping the United States and the world. Their website includes a question search tool that returns results from a database of survey questions on various topics (http://www.pewresearch.org/question-search/). We used the tool to find survey questions that contain unfamiliar wordings which would be relevant for our experiment. The starting point was a list of 60 words that we randomly selected from the list of wordings that have a frequency between 800 and 1000 in the BNC. As we explained earlier (in Section 4.1.1), the BNC was used because of the availability of the downloadable database of words.

For each wording, we manually searched for synonyms and other similar wordings in the WordNet lexical database, but only 31 of the 60 wordings had at least one alternative with a higher frequency than the original. We then looked up these 31 wordings in the PEW database and found that 11 of them (justified, relate, notify, luxury, physician, prone, compact, mandatory, enjoyable, nationally and civic) were used in at least one PEW question. We not only looked at single questions but complete questionnaires that contained these words as well.

Among them, we arbitrarily selected a questionnaire on the topic of terrorism that contained the wordings ‘justified’ and ‘prone’. We searched for other questions on the topic of terrorism and came to a final selection of eight questions (or 12 items) in the following order:

P1. In general, how well do you think the United States government is doing in reducing the threat of terrorism?
P2. How worried are you that there will soon be another terrorist attack in the United States?
P3. Do you think the use of torture against suspected terrorists in order to gain important information can ever be justified?

P4. Do you completely agree, mostly agree, mostly disagree, or completely disagree with this statement?

P4.1. I often worry about the chances of a nuclear attack by terrorists.

P4.2. Freedom of speech should not extend to groups that are sympathetic to terrorists.

P4.3. The police should be allowed to search the houses of people who might be sympathetic towards terrorists without a court order.

P4.4. The government's anti-terrorism policies have gone too far in restricting the average person's civil liberties.

P4.5. I am concerned that the government is collecting too much information about people like me.

P5. As you may know, the United States government has a policy that it NEVER pays ransom money for hostages held by terrorist groups. Overall, do you approve or disapprove of this policy?

P6. Which statement comes closer to your own views even if neither is exactly right? Please select:

- Some religions are more prone to violence than others.
- All religions are about the same when it comes to violence.

P7. Which statement comes closer to your own views even if neither is exactly right? Please select:

- The Islamic religion is more likely to encourage violence among its believers.
- The Islamic religion does not encourage violence more than others.

P8. How concerned, if at all, are you about Islamic extremism around the world these days?

As for the Wageindicator case in the previous subsection, we looked up each of the underlined wordings in WordNet online (Table 4.2): We looked for synonyms in the synset, other similar words, and hyponyms and hypernyms. In addition, for some wordings – those for which we did not find enough suitable alternatives – we also considered words used in the definitions (e.g., 'disposed to' for 'sympathetic') and other synsets in the search query (other meanings and other parts of speech). For one of the wordings, we also looked in the Microsoft Office thesaurus for additional alternatives.

Table 4.2: Synonyms and other alternative wordings found in WordNet: PEW questions

Case	Query	WordNet Synset	Definition
P1	threat	**menace, threat** (n)	something that is a source of danger
	hyponym	yellow peril (n)	the threat to Western civilization said to arise from the power of Asiatic peoples
	hypernym	**danger** (n)	a cause of pain or injury or loss
P2	worried	**apprehensive, worried** (adj)	mentally upset over a possible misfortune or danger, etc.
	similar to	**uneasy** (adj)	lacking sense of security or affording no ease or reassurance
	other meaning	disquieted, distressed, disturbed, **upset, worried** (adj)	afflicted with or marked by anxious uneasiness or trouble or grief
	similar to	troubled (adj)	characterized by or indicative of distressed or affliction or danger or need
	other PoS	**concern**, interest, occupy, worry (v)	be on the mind of
P3	justified	**justify, vindicate** (v)	show to be right by providing justification or proof
	troponym	**excuse**, explain (v)	serve as a reason or cause or justification of
	troponym	**legitimate** (v)	show or affirm to be just and legitimate
	troponym	**warrant** (v)	provide adequate grounds to justify (a certain course of action)
	hypernym	uphold, maintain (v)	support against an opponent
P4.1	chances	**probability, chance** (n)	a measure of how likely it is that some event will occur, a number expressing the ratio of favorable cases to the whole number of cases possible
	hyponym	**risk**, risk of exposure (n)	the probability of being exposed to an infectious agent
	hyponym	… [7 other hyponyms] (n)	… [various definitions]
	hypernym	measure, quantitiy, amount (n)	how much there is or how many there are of something that you can quantify
P4.2 P4.3	sympathetic to	**sympathetic** (adj)	expressing or feeling or resulting from sympathy or compassion or friendly fellow feelings, disposed towards
	definition	**disposed to**	naturally disposed toward
	see also	**compassionate** (adj)	showing or having compassion
	see also	congenial (adj)	suitable to your needs
	see also	**kind** (adj)	having or showing a tender or considerate and helpful nature
	similar to	commiserative (adj)	feeling or expressing sympathy
	similar to	condolent (adj)	expressing sympathy with a person who experienced the death of a loved one
	similar to	empathic, empathetic (adj)	showing empathy or ready comprehension
	other PoS	feel for, pity, compassionate, condole with, **sympathize with** (v)	share the suffering of
	MS Word thesaurus	**favor**, favour (n)	an inclination to approve
	MS Word thesaurus	**supportive** (adj)	furnishing support or assistance
	other PoS	**support**, back up (v)	give moral or psychological support, aid, or courage to

Case	Query	WordNet Synset	Definition
P4.4	restricting	**restrict, curtail, curb, cut back (v)**	place restrictions on
	troponym	**abridge** (v)	lessen, diminish, or curtail
	troponym	immobilize, immobilise (v)	cause to be unable tomove
	troponym	ration (v)	restrict the consumption of a relatively scarce commodity, as during war
	troponym	restrict, **control** (v)	place under restrictions; limit access to by law
	hypernym	**limit**, circumscribe, confine to (v)	restrict or confine within limits
P4.5	collecting	**gather, garner, collect, pull together** (v)	assemble or get together
	definition	**assembling**	collect in one place
	troponym	… [19 other troponyms]	… [various definitions]
P5a	ransom money	**ransom, ransom money**	money demanded for the return of a captured person
	definition	**demanded** (v)	request urgently and forcefully
	hypernym	cost (n)	the total spent for goods or services including money and time and labor
P5b	hostages	**hostage, surety** (n)	a prisoner who is held by one party to insure that another party will meet specified terms
	hypernym	prisoner, captive (n)	a person who is confine; especially a prisoner of war
P6	prone	**prone** (adj)	having a tendency (to)
	similar to	**inclined** (adj)	having a preference, disposition, or tendency
P7	encourage	**promote, advance, boost, further, encourage** (v)	contribute to the progress or growth of
	troponym	… [7 other troponyms]	… [various definitions]
	hypernym	support, back up (v)	give moral or psychological support, aid, or courage to
P8	concerned	**concerned** (adj)	feeling or showing worry or solicitude
	see also	attentive (adj)	giving care or attention
	see also	**troubled** (adj)	characterized by or indicative of distressed or affliction or danger or need
	similar to	afraid (adj)	filled with regret or concern; used often to soften an unpleasant statement
	similar to	**afraid** (adj)	filling worry or concern or insecurity
	similar to	haunted, obsessed, **preoccupied**, taken up (adj)	having or showing excessive or compulsive concern with something
	similar to	**solicitous** (adj)	full of anxiety and concern
	other PoS	**concern**, interest, occupy, **worry** (v)	be on the mind of

*Note: The column 'Query' represents the word that we entered in the WordNet browser (available at http://wordnetweb.princeton.edu/perl/webwn). An example of query results for case W1 is presented in the Appendix (see Figure B.2). Moreover, it is also possible to click and open additional layers to retrieve also hyponyms, hypernyms and other alternative wordings (see Figure B.3 in the Appendix).

The selected alternative wordings (bolded) will be examined with the COCA and enTenTen corpus and then, in Sections 4.3.3 and 4.3.4, with expert reviews. For the

noun ‘threat’ (P1), we selected the synonym ‘menace’ and the hypernym ‘danger’. For the adjective ‘worried’ (P2), we found two suitable senses: First, for ‘mentally upset over a possible misfortune’, we selected the synonym ‘apprehensive’ and the similar word ‘uneasy;’ second, for ‘afflicted anxious uneasiness’, we selected the synonym ‘upset’ and the similar word ‘troubled’. In addition, we also considered its verbal form, where a suitable synonym is ‘concern’ but the suitable form is its past participle ‘concerned’. For the verb ‘justified’ (P3), we used the synonym ‘vindicate’ and three troponyms: ‘excuse’, ‘legitimate’ and ‘warrant’. For the noun ‘chances’ (P4), we used only the synonym ‘probability.’ The most difficult and complex wording in this case study is the adjective ‘sympathetic to’ (P4.2 and P4.3), for which we first took one alternative from the definition (‘disposed to’), two similar wordings (‘compassionate’ and ‘kind’) and the verbal form ‘sympathise with’. We considered the list of available alternatives for this case insufficient and decided to look further: We checked the synonymous wordings in the Microsoft Word thesaurus, where there are several other alternatives – we took the noun ‘favour’ (in favour) and the adjective ‘supportive’, for which we also considered its verbal form, ‘support’. For the verb ‘restricting’ (P4.4), we selected the three synonyms (‘curtail’, ‘curb’, and ‘cut back’), two of the several troponyms (‘abridge’ and ‘control’), and one hypernym (‘limit’). For the verb ‘collecting’ (P4.5), we selected three synonyms (‘gather’, ‘garner’ and ‘pull together’) and the verb used in the definition: ‘assemble’. For ‘ransom money’ (P5a), we selected the synonym (‘ransom’) and the verb from the definition (‘demanded’). For the noun ‘hostages’ (P5b), we selected the synonym ‘sureties’. For the adjective ‘prone’ (P6), we selected the similar word ‘inclined’. For the verb ‘encourage’ (P7), we selected four synonyms: ‘promote’, ‘advance’, ‘boost’, and ‘further’. For the adjective ‘concerned’ (P8), we selected four similar wordings (‘troubled’, ‘afraid’, ‘preoccupied’ and ‘solicitous’), and we formed the verbal form ‘worried’ based on the relation between the verbs ‘concern’ and ‘worry’.

4.2 Evaluations based on the text corpora approach

To some extent, corpora were used already in the selection of question items (Section 4.1); however, that was only preliminary and limited to one corpus (BNC). In this section, we fully elaborate the corpora approach by computing wording frequencies, both single and multi-word strings, using two additional corpora that enable the retrieval

of this kind of information: COCA and enTenTen. In fact, different corpora are composed differently and might give different results Moreover, it is not enough to check only single frequencies – the context of the sentence is also important and word frequencies of strings of words need to be retrieved.

It should be noted that although multiple corpora were used, their absolute frequencies were never directly compared against each other: only the resulting order relations were compared. In fact, each corpus has a different size and to allow such comparisons, we would need to use normalised frequencies (e.g., for millions). Thus, we decided to limit our focus by only comparing frequencies within the same corpus – that is, if a frequency of a certain word was less than or greater than the frequency of its synonym or other alternative word. When two different corpora were compared, we limited our focus only to order relations.

4.2.1 The Wageindicator case

We looked up frequencies for the specific context for all the wordings in bold in Table 4.1, both for single words and strings, as presented in Table 4.3.

Table 4.3: Wording frequencies based on enTenTen corpora (Wageindicator questionnaire)

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
W1	form	89988	4834526	form of contract	4	872
	kind	185404	4149346	kind of contract	15	954
	sort	94356	1941451	sort of contract	14	471
	type	50215	5073343	type of contract	15	2092
	variety	37626	1922555	variety of contract	2	229
W2	all	14405156	34568083	all owned	27	927
	completely	37697	1548184	completely owned	4	287
	entirely	24939	591514	entirely owned	6	201
	totally	24191	868971	totally owned	5	214
	wholly	3550	103214	wholly owned	3	16383
W3	established	30421	1707542	established in the contract	1	38
	laid down	1331	37282	laid down in the contract	0	16
	made	387626	7911750	made in the contract	0	48
	set	183469	6349179	set in the contract	0	40
W4	out of the blue	1045	18618	out of the blue problems	0	7
	unanticipated	767	14409	unanticipated problems	10	328
	unexpected	9218	230846	unexpected problems	32	1278

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
	unforeseen	797	36496	unforeseen problems	23	1004
	unlooked-for	0	428	unlooked-for problems	0	0
	unseen	2946	45485	unseen problems	1	100
W5.1	adequate	11835	362254	adequate energy	13	577
W5.2				adequate support	78	2383
W8				staffing levels are adequate	0	8
	enough	172956	3626045	enough energy	338	9238
				enough support	181	3894
				there is enough staff	0	7
	sufficient	11609	495677	sufficient energy	26	1645
				sufficient support	46	1336
				staffing levels are sufficient	0	3
W6	adequately	4589	111679	adequately varied	0	1
	sufficiently	4685	107690	sufficiently varied	4	55
				varied enough	24	419
W7	condition	28624	2789935	good condition	517	53450
	state of repair	19	1280	good state of repair	2	258

Since most of the COCA string frequencies are zero, we are only going to interpret the enTenTen frequencies.

The original wording, the adjective ‘kind’ (W1), has the highest wording in the COCA corpus but is only third (after ‘type’ and ‘form’) according to the enTenTen corpus. Moreover, looking at the broader context, ‘kind of employment contract’ (7) is less frequent than ‘type of employment contract’ (28) and ‘form of employment contract’ (11). However, the differences are quite small and further evaluation by experts is needed.

Instead of analysing two different wording contexts, ‘wholly domestic-owned’ and ‘wholly foreign-owned’, which both have very low frequencies in text corpora, we focus on their common denominator, ‘wholly owned’ (W2), which has a higher frequency. If considering only the single frequency, the adverb ‘wholly’ is the least frequent choice among the alternatives; however, the wording ‘wholly owned’ has the highest frequency in the corpus (16,383) and the alternatives are all lower: ‘all owned’ (927), ‘completely owned’ (287), ‘entirely owned’ (201) and ‘totally owned’ (214).

The original verb conjugation ‘laid down’ (W3) has the lowest frequency, both as a single word and in the context ‘laid down in the contract’ (16). The best alternative is ‘made in the contract’ (48), followed by ‘set in the contract’ (40) and ‘established in the contract’ (38).

The wording ‘unforeseen problems’ (W4) has only the third highest single frequency and second highest string frequency (1,004). The highest frequency, both single and contextual, was observed for ‘unexpected problems’ (1,278). Other alternatives have lower contextual frequencies: ‘unanticipated problems’ (328), ‘unseen problems’ (100) and ‘out of the blue problems’ (7). The least frequent is ‘unlooked-for problems’, which does not appear in the corpora.

The adjective ‘enough’ not only has the highest single frequency compared to ‘adequate’ and ‘sufficient’, but also the highest frequency in three contexts: ‘energy’ (W5.1), ‘support’ (W5.2) and ‘varied’ (W8). ‘Enough energy’ (9,238) has a higher frequency than both the original, ‘sufficient energy’ (1,645), and ‘adequate energy’ (577). ‘Enough support’ (3,894) has a higher frequency than ‘adequate support’ (2,383) and the original, ‘sufficient support’ (1,336). Also, in the adverbial context, ‘varied enough’ (W6) has a higher frequency (419) than ‘sufficiently varied’ (55) and ‘adequately varied’ (1). There is a fourth context, ‘staffing levels’ (W8), where all the frequencies are very low. Thus, it does not make much sense to compare the frequencies: ‘there is enough staff’ (7) is about the same as ‘staffing levels are adequate’ (8), and both have a slightly higher frequency than the original ‘staffing levels are sufficient’ (3).

The wording ‘good conditions’ (W7) has a much higher wording frequency (53,450) than ‘good state of repair’, both single and for the string.

4.2.2 The PEW case

Wording frequencies for words in bold in Table 4.2 were retrieved from the COCA and enTenTen corpus, both for single words and strings of words, as presented in Table 4.4.

Table 4.4: Wording frequencies based on COCA and enTenTen corpora (PEW questions)

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
P1	danger	20370	512584	danger of terrorism	8	116
	menace	1866	50138	menace of terrorism	2	126
	threat	30382	666925	threat of terrorism	192	2209
P2	apprehensive	920	18993	how apprehensive	4	30
	concerned	39502	776023	how concerned	114	963
	uneasy	3386	35863	how uneasy	6	104
	upset	15417	265608	how upset	122	1696
	worried	25024	324153	how worried	181	824
P3	excused	1272	17263	ever excused	0	8
	legitimate	10844	279690	ever legitimate	3	33
	justified	5038	111269	ever justified	10	227
	vindicated	742	8863	ever vindicated	1	3
	warranted	1595	36895	ever warranted	2	31
P4.1	chances	12915	482356	chances of attack	1	49
	probability	5075	222141	probability of attack	2	28
	risk	64294	2526058	risk of attack	15	483
P4.2-3	compassionate to	40	1022	compassionate to terrorists	0	0
	disposed to	352	7501	disposed to terrorists	0	0
	in favour of	213	65112	in favour of terrorists	0	1
	kind to	1251	38483	kind to terrorists	0	2
	support	120828	6192586	support terrorists	18	458
	supportive of	1948	29212	supportive of terrorists	1	3
	sympathetic to	1319	12456	sympathetic to terrorists	1	18
	sympathize with	712	13030	sympathize with terrorists	0	16
P4.4	abridging	56	906	abridging liberties	0	1
	controlling	4	2179	controlling liberties	0	0
	curbing	610	12729	curbing liberties	0	1
	curtailing	319	4859	curtailing liberties	0	6
	cutting back	1000	15578	cutting back liberties	0	0
	limiting	4514	116859	limiting liberties	1	4
	restricting	1665	39914	restricting liberties	1	7
P4.5	assembling	1481	32420	assembling information	7	83
	collecting	7573	208996	collecting information	154	4417
	garnering	308	10618	garnering information	1	37
	gathering	11133	340008	gathering information	312	7542
	pulling together	199	4261	pulling together information	0	48
P5a	demanded money	41	856	demanded money for hostages	0	0
	ransom	1330	27098	ransom for hostages	5	9
	ransom money	41	506	ransom money for hostages	0	1

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
P5b	hostages	4627	15729	money for hostages	1	1
	sureties	7	1570	money for surities	0	0
P6	inclined	4716	118399	inclined to violence	3	39
	prone	3884	140703	prone to violence	59	452
P7	advance	17699	1438599	advance violence	0	4
	boost	9625	603199	boost violence	0	2
	encourage	17136	1311056	encourage violence	9	517
	further	64650	2472985	further violence	69	922
	promote	15942	1488081	promote violence	29	1002
P8	afraid	31099	411099	afraid about extremism	0	0
	concerned	39502	776023	concerned about extremism	1	3
	preoccupied	2148	21312	preoccupied about extremism	0	0
	solicitous	335	2327	solicitious about extremism	0	0
	troubled	8576	111411	troubled about extremism	0	0
	worried	25024	324153	worried about extremism	0	0

The original noun ‘threat’ (P1) has a much higher frequency than ‘danger’ and ‘menace’, both in the COCA and enTenTen. This is probably due to the newspaper genre, which is strongly represented in corpora. Also, the context ‘threat of terrorism’ (2,209) is the alternative with the highest frequency, being much higher than ‘menace of terrorism’ (126) and ‘danger of terrorism’ (116).

Considering only single frequencies, the adjective ‘concerned’ (P2) is the best choice according to both corpora, followed by the original wording ‘worried’ and then ‘upset’. However, the collocation with the highest frequency is ‘how upset’ (1,696), probably because it is too broad and has several senses. Apparently, being limited to the combination with the adverb ‘how’ is not much better than studying only the single wordings, and extending the context to include more words should be considered. However, longer strings of words, such as ‘worried that there will be an attack’ and other alternatives and combinations, return zero frequencies in the corpora. Thus, we analyse word sequences of size two: ‘how concerned’ (963) has a slightly higher frequency than the original wording ‘how worried’ (824). On the other hand, ‘how apprehensive’ (30) and ‘how uneasy’ (104) have the lowest frequencies.

The adjective ‘legitimate’ (P3) has the highest frequency in both corpora, followed by the original wording ‘justified’. However, the string ‘ever justified’ has the highest word frequency (227), which is much higher than the other alternatives: ‘ever legitimate’ (33), ‘ever warranted’ (31), ‘ever excused’ (8) and ‘ever vindicated’ (3).

The noun ‘risk’ (P4.1) has the highest frequency in both corpora, and the string ‘risk of attack’ is the wording alternative with the higher frequency (483), while the original wording ‘chances of attack’ is in second place (49). ‘Probability of attack’ has the lowest frequency (28).

The noun ‘support’ (P4.2 and P4.3) is the single wording with the highest frequency among the alternatives, and ‘support terrorists’ is the wording with the highest frequency (458). The original wording ‘be sympathetic to terrorists’ has a much lower frequency (18), which is about the same as ‘sympathise with terrorists’ (16) and not much higher than ‘be supportive of terrorists’ (3), ‘kind to terrorists’ (2) and ‘in favour of terrorists’ (1). On the other hand, the wordings ‘compassionate to terrorists’ and ‘disposed to terrorists’ do not appear in any of the corpora.

The verb ‘limiting’ (P4.4) would be the best alternative considering only single wordings. Moreover, the original wording ‘restricting liberties’, has only a slightly higher frequency (7) than ‘curtailing liberties’ (6) and ‘limiting liberties’ (4). ‘Abridging liberties’ (1), ‘controlling liberties’ (0), ‘curbing liberties’ (1) and ‘cutting back liberties’ all have lower enTenTen frequencies and are considered inappropriate by experts.

The verb with the higher frequency, both single and string, is the wording ‘gathering information’ (7,542) (P4.5), while the original wording ‘collecting information’ has only the second highest frequency (4,417). The third choice is ‘assembling information’ (83), the fourth is ‘pulling together information’ (48), and the least frequent is ‘garnering information’ (37).

Using only ‘ransom for hostages’ (P5a and P5b) gives a slightly higher frequency (9) than the original ‘ransom money for hostages’ (1), while the alternative ‘demanded money for hostages’ does not appear in the corpora. ‘Hostages’ is more frequent than the synonym ‘sureties’; in particular, ‘money for sureties’ does not even appear in the corpora.

As the originally used adjective ‘prone’ (P6) has a higher frequency in both corpora than ‘inclined’, the wording ‘prone to violence’ is much more frequent (452) than the alternative wording ‘inclined to violence’ (39).

While the verb ‘further’ (P7) would be the best choice considering only single frequencies, it is surpassed by ‘promote’ when considering the context. In fact, ‘promote violence’ has the highest frequency (1,002), followed by ‘further violence’ (922). The original wording ‘encourage violence’ has a notably lower (517) wording than the first two, and ‘advance violence’ (4) and ‘boost violence’ (2) have even lower frequencies.

The original verb ‘concerned’ (P8) has the highest frequency in both corpora and the string ‘concerned about extremism’, as it is the only wording with a frequency higher than zero (3); while ‘worried about extremism’, ‘troubled about extremism’, ‘afraid about extremism’, ‘preoccupied about extremism’ and ‘solicitous about extremism’ have zero appearances in the corpora.

4.3 Expert evaluations

We now present the alternative to the text corpora approach, which is based on expert evaluations. The experts were exposed to the same alternatives as those compared with linguistic resources. First, we present the instrument that was used in the evaluations (Section 4.3.1). Second, we present how data was collected, i.e., who the experts are and what are their characteristics (Section 4.3.2). Third, we present the results for the Wagendicator (Section 4.3.3) and PEW case (Section 4.3.4). Finally, we present some criticism of the methodology that was given by experts (Section 4.3.5) and compare the results of evaluations of native and non-native speaking experts (Section 4.3.6).

4.3.1 Methodological approach

We invited 132 experts to evaluate either the nine Wageindicator items or the 12 PEW items by completing a semi-structured online questionnaire. In total, 81 experts responded to our invitation: 17 of them evaluated both studies, while the remainder evaluated only one study. In the following subsection, we first present the questionnaire that was used to measure the subjective appropriateness of different wordings, the

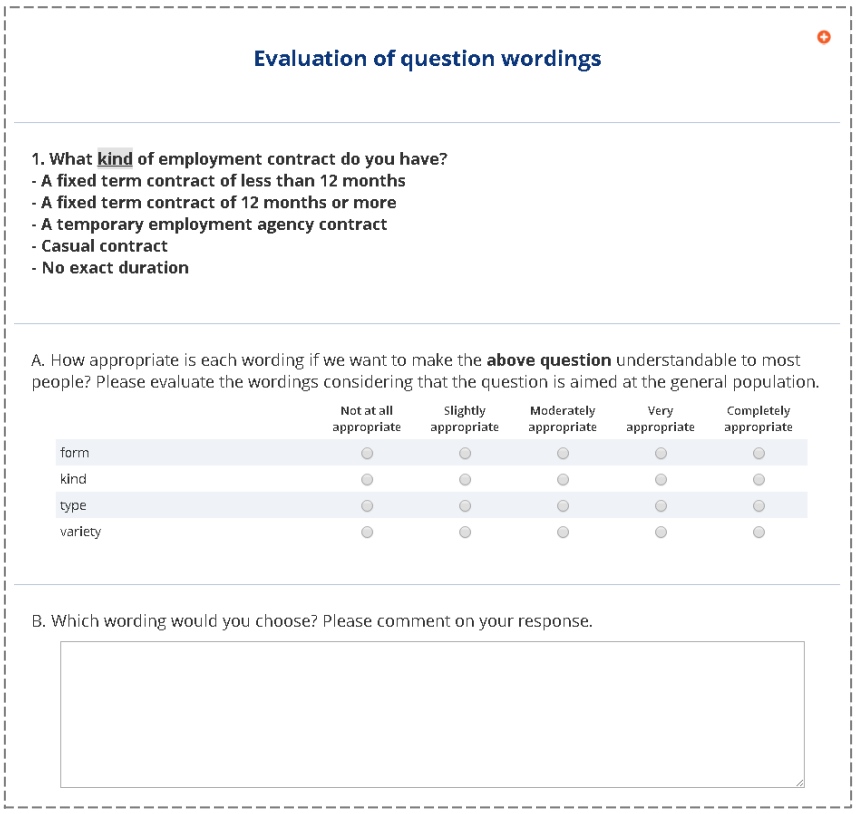
experts' preferences and other comments, and then also details about how data were collected (Section 4.3.2).

The first screen of the evaluation questionnaire contained instructions describing the required task (Figure B.5 in Appendix), that is:

A. Evaluating the appropriateness of different wordings (by which we mean the wording that makes the question meaning clearest and easiest to understand for the general population).
B. Indicating the preferred wording and explaining the choice.

The items to be evaluated then followed, each on its own screen, accompanied by two probing questions. The wording that was supposed to be evaluated in each item was underlined, and the alternative wordings were displayed on mouse-over in addition to being listed within probing Question A. In the Wageindicator evaluation, the evaluated wording (underlined) was always the word originally used in the master questionnaire (displayed in Section 4.1.1), as it was usually the word with the lower frequency. Figure 4.1 presents an example for the wording 'kind' (W1).

Figure 4.1: Example of evaluation question



The wording that was supposed to be evaluated in each item was underlined and the alternative wordings were displayed on mouseover, in addition to being listed also within probing question A. In the Wageindicator evaluation, the evaluated wording (underlined) was always the word originally used in the master questionnaire (displayed in Section 4.1.1) as that was usually the word with the lower frequency.

In the PEW case, on the other hand, we did not always use the original wording from the master questionnaire, but decided to choose one of the wordings that had a lower wording frequency than the original, based on Table 4.4. The decision was based on the assumption that an expert would tend to prefer the displayed, underlined wording. Given that we also assumed they would better evaluate the higher-frequency wording, we wanted to avoid a confounding effect by showing the wording that was less frequent. Thus, we changed the original wording in the master questionnaire (as displayed in Section 4.1.2) with a lower-frequency wording (as displayed in Table 4.5) in all cases, except for 'ransom money' (P5a) and 'prone' (P6).

Table 4.5: Wording displayed in evaluation questionnaire

	Original PEW wording	Displayed wording
P1	threat	menace
P2	worried	apprehensive
P3	justified	vindicated
P4.1	chances	probability
P4.2-3	sympathetic to	disposed to/towards
P4.4	restricting	curtailing
P4.5	collecting	assembling
P5a	ransom money	ransom money
P5b	hostages	sureties
P6	prone	prone
P7	encourage	boost
P8	concerned	preoccupied

4.3.2 Data collection

The list of 132 experts was generated based on the authors' contacts with researchers in the field of survey methodology and other relevant fields. The experts included contacts that we made within various scientific communities, in particular the Webdatanet network (http://www.webdatanet.eu/), attendants of the Internet Survey Workshop (http://workshop.websm.org/), members of the European Survey Research Association

(http://www.europeansurveyresearch.org/), and members of the American Association for Public Opinion Research (https://www.aapor.org/). To be considered an expert, one had to have at least some expertise developing survey questionnaires; however, we mostly focused on those who had more expertise, primarily those who presented at questionnaire design sessions at recent conferences. The experts came from various countries, mostly in Europe: 17% from the UK and 55% from other European countries (Austria, Belgium, Croatia, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Iceland, Italy, Luxembourg, Macedonia, the Netherlands, Norway, Poland, Portugal, Romania, Russia, Spain and Switzerland). Of the remaining experts, 18% came from the US and 10% came from other non-European countries (Australia, Canada, Israel, Japan, Mexico, Peru and Turkey)

Half of the experts were allocated to the Wageindicator questions and the other half to the PEW questions (1st wave). Invitations were sent by e-mail (Figure B.4 in Appendix) in May 2015 (1st evaluation) and June and July 2015 (2nd evaluation). The allocation was not random but arbitrary – members of the Webdatanet network and attendants of the Internet Survey Workshop were mostly included in the Wageindicator evaluation because they were familiar with the study, while the members of the two professional societies were mostly included in the PEW question evaluation. When an expert was a member of more than one group, he or she was randomly allocated to the remaining places on either list. Thus, the country structure of those invited to respond to the WageIndicator evaluation was more European: 22% were from the UK and 73% were from other European countries, while only 5% were from the US and 8% were from other non-European countries. On the other hand, among the experts invited to evaluate PEW questions, 31% were from the US and 13% were from other non-European countries, while only 12% were from the UK and 56% were from other European countries. The different structure was intentional because we wanted to have more Europeans for the WageIndicator questionnaire, which is written in British English, and less Europeans for the PEW questions, which are written in American English.

In addition, some of the experts told us that they forwarded the invitation to participate to other experts who were not on our list. To our knowledge, the first evaluation was sent to seven additional experts and the second evaluation to six additional experts. Moreover, at the end of both evaluations, the experts were asked if they wanted to

participate in another evaluation – 10 respondents from the first survey and 17 respondents from the second survey agreed to participate in another evaluation. We sent them an additional invitation a few weeks after they had participated in the first evaluation (2nd wave): in June, July and August. Thus, in total, there were at least 89 invitations sent for the first questionnaire evaluation, and at least 83 for the second.

In total, 76 experts started responding to the first evaluation, but only 51 made it to the end, among them six out of the 10 respondents who responded to the second evaluation first. On the other hand, 79 experts started responding to the second evaluation, but only 55 made it to the end, among them 11 of the 17 respondents who also responded to the first evaluation. However, not all responses were complete: We excluded two incomplete evaluations of the first questionnaire and six incomplete evaluations of the second. We thus analysed 49 evaluations of the first questionnaire and 49 evaluations of the second questionnaire (Table 4.6).

Table 4.6: Number of invitations sent and response rate

	First evaluation (Wageindicator)	Second evaluation (PEW)
Invitations sent – 1st wave	66	66
Invitations forwarded (estimate)	6	7
Invitations sent – 2nd wave	17	10
Invitations total	89	83
Started responding	76 (85% of invited)	79 (95% of invited)
Made it to the end of the evaluation	51 (67% of those who started)	55 (70% of those who started)
Complete responses	49	49
Agreed to participate in 2nd wave	10 (20% of respondents)	17 (33% of respondents)
Participated in both evaluations	6 (60% of those who agreed)	11 (65% of those who agreed)

Respondents took on average eight minutes and 40 seconds to respond to the Wageindicator evaluation and 13 minutes and 28 seconds to respond to the PEW question evaluation, which contained more question items. The results are presented in Sections 4.3.3 and 4.3.4.

4.3.3 Results for the Wageindicator case

First, we present the results of the first question in the Wageindicator questionnaire evaluation to which 49 experts responded, out of which 18 are native English speakers and an additional 12 have lived in an English-speaking country for at least one year.

About half of them work in academia (26), while the others work in industry (11), government (9), or have other affiliations (non-profit, own survey business). Most of the responding experts (31) took graduate-level courses in questionnaire design; some also took graduate-level courses in cognitive science (20) and a few also in linguistics (6). Most of the experts indicated survey methodology or questionnaire design as their main area of expertise, but some also indicated other fields, including statistics, clinical psychology, engineering, labour economics, design of forms, social research, business economy and management, sociology, political science, market research and applied linguistics and demography. We also asked the experts to self-evaluate their own expertise regarding the development of survey questionnaires: 14 labelled themselves as experts, 22 as very experienced, eight as somewhat experienced, and four as having little experience; none, however, responded as having no experience.

Table 4.7 presents responses to the question, 'How appropriate is each wording if we want to make the corresponding question understandable to most people?' (Question A in Figure 4.1). Shaded cells are those with a frequency over 20%, and the best wording option according to experts for each item is also shaded.

Table 4.7: Appropriateness of wordings in Wageindicator study

Wording	Completely appropriate	Very appropriate	Moderately appropriate	Slightly appropriate	Not at all appropriate	n
form	2.0%	12.2%	38.3%	30.6%	16.3%	49
kind	35.4%	41.7%	14.6%	6.3%	2.1%	48
type	54.2%	35.4%	10.4%	0.0%	0.0%	48
variety	0.0%	2.1%	12.5%	25.0%	60.4%	48
all	2.2%	4.3%	15.2%	28.3%	50.0%	46
completely	37.5%	35.4%	16.7%	6.3%	4.2%	48
entirely	37.5%	37.5%	18.8%	6.3%	0.0%	48
totally	16.7%	31.3%	31.3%	12.5%	8.3%	48
wholly	17.0%	19.1%	25.5%	21.3%	17.0%	47
established	8.5%	21.3%	29.8%	21.3%	19.1%	47
laid down	6.4%	12.8%	31.9%	27.7%	21.3%	47
made	2.1%	0.0%	10.6%	25.5%	61.7%	47
set	36.7%	28.6%	26.5%	6.1%	2.0%	49
out of the blue	0.0%	2.2%	21.7%	26.1%	50.0%	46
unanticipated	20.8%	37.5%	35.4%	2.1%	4.2%	48
unexpected	55.1%	44.9%	0.0%	0.0%	0.0%	49
unforeseen	26.5%	38.8%	30.6%	2.0%	2.0%	49
unlooked-for	0.0%	2.1%	8.5%	40.4%	48.9%	47

Wording	Completely appropriate	Very appropriate	Moderately appropriate	Slightly appropriate	Not at all appropriate	n
unseen	0.0%	6.4%	6.4%	21.3%	66.0%	47
adequate	10.9%	32.6%	28.3%	15.2%	13.0%	46
enough	42.6%	38.3%	12.8%	4.3%	2.1%	47
sufficient	19.6%	56.5%	19.6%	2.2%	2.2%	46
adequately varied	8.3%	14.6%	35.4%	29.2%	12.5%	48
sufficiently varied	22.4%	28.6%	32.7%	12.2%	4.1%	49
varied enough	27.1%	16.7%	22.9%	20.8%	12.5%	48
conditions	38.8%	32.7%	6.1%	10.2%	12.2%	49
state of repair	23.9%	17.4%	32.6%	13.0%	13.0%	46
Staffing levels are adequate.	20.4%	28.6%	28.6%	14.3%	8.2%	49
Staffing levels are sufficient.	14.6%	31.3%	43.8%	4.2%	6.3%	48
There is enough staff.	31.3%	35.4%	20.8%	4.2%	8.3%	48

Note: Shaded all cells over 20%

A certain divergence between corpora frequencies (Table 4.3) and expert evaluations (Table 4.7) can be observed. In the following subsections, we analyse and compare the results for each individual wording item separately, including a detailed analysis of open-ended responses (Question B in Figure 4.1). We coded experts' comments into 13 categories, where odd numbers represent disadvantages and even numbers represent advantages of respective wordings (Table 4.8). Each wording can have several codes, and we use them in the following sections to label how a certain response was coded.

Table 4.8: Coding categories for the open question

Code	Description	Code	Description
1	Difficult to understand	2	Easy to understand
3	Uncommon word	4	Commonly used word
5	Uncommon in context	6	Commonly used in context
7	Vague meaning	8	Univocal
9	Wrong connocation	10	Right connotation
11	Gramatically incorrect	12	Other pros
13	Other cons		

4.3.3.1 Kind of employment contract (W1)

'Type of contract' is the most frequent wording in the enTenTen corpus (2,092) and is also the most appropriate according to experts: 54% consider it completely appropriate

and 35% very appropriate. The second preferred choice among experts is the wording ‘kind of employment contract’ (35% completely appropriate and 42% very appropriate), which also has the second largest frequency (954) in the enTenTen corpus. It is also the original wording used in the Wageindicator questionnaire. On the other hand, the wording ‘form of employment contract’, with a frequency of 872, is only moderately (38%) or slightly (31%) appropriate according to most experts; while the worst choice, according to both experts and the corpora, is ‘variety of employment contract’, which has a frequency of 229 in the corpus and is considered not at all appropriate by 60% of experts and only slightly appropriate by an additional 25%.

Twenty-one out of 49 experts chose ‘type’, for which they gave various reasons. Two experts wrote that it is the ‘easiest to understand’ or ‘most familiar to respondents’ (Code 2). Five experts commented that it ‘fits more with the question’, as we often talk about ‘types of things’, or that it is a ‘standard word’ – or, more simply, that it is a ‘commonly used phrase in this context’. One expert even mentioned ‘type’ as being ‘often used in survey questionnaires’ (Code 6). Three experts argued that it is the ‘least ambiguous’, ‘more precise’ or ‘more concrete’ wording (Code 8). Finally, six experts thought that it has the most appropriate meaning. Two of them believed that it is more formal than ‘kind’, while the others explained what they thought it means, i.e., ‘defines a category or group’, ‘implies there are different versions or variations’, ‘suggests a category of things with common characteristics’, or ‘describes best that you are interested in the duration of the employment contract’ (Code 10).

The experts also provided explanations for why the other terms are less appropriate or not at all appropriate. ‘Form’ is less commonly used (Code 3), ‘can be understood in two different ways’ (Code 7), ‘is associated with paper form’, wrote one expert, and ‘might make some respondents think of the actual paper format of the contract’, explained another (Code 9). Although ‘kind’ is ‘more common in spoken language’ (Code 4) and is a collocate (Code 6), it does not fit as well with the question as ‘form’ (Code 5), is less ‘concrete’ as it also ‘has another meaning (‘nice’ or ‘gentle’), and ‘it is generally better to choose terms that have a single meaning’ (Code 7). Moreover, according to four experts, it sounds a bit too informal and colloquial (Code 9). Finally, two experts argued that ‘variety’ would usually not be used in this context (Code 5), and another argued that it is ‘just the wrong word, doesn’t make much sense’ (Code 9).

On the other hand, 17 experts chose ‘kind’, with three of them arguing that it is simple and easy to understand (Code 2). Two of the experts wrote that it is more commonly used (Code 4), and three specified that it is common in the context, one of whom mentioned that it is commonly heard in the media (Code 6). Moreover, one expert noted that it is ‘less formal’, and three experts noted that it conveys the correct meaning, one of whom explained that ‘type refers to a specific theoretical contract and kind to a practical contract’ (Code 9). Most experts did not provide any explanation for why ‘type’ is less appropriate, except for two – one wrote that it made him ‘think of formalities’ and another that ‘type would address to full-time or part-time’ (Code 9).

Six experts could not decide between ‘kind’ and ‘type’ but did not explain why, while one was also undecided about ‘form’. Finally, two experts suggested rewording the question so that it would not include any of the proposed words, i.e., ‘Which of the following best describes your employment contract?’ or ‘Which type of employment?’

To sum up, according to both the corpora and the experts, the best choice is ‘type of employment’ contract. Even those experts who did not put it as their first choice considered it at least slightly appropriate, and only two of them indicated that it might have some wrong connotations. Therefore, we recommend replacing ‘kind of contract’ – the wording used in the Wageindicator questionnaire – with ‘type of contract’.

4.3.3.2 Wholly owned (W2)

As mentioned in Section 4.2.1, we decided to analyse the more frequent phrase ‘wholly owned’ as the wording frequencies of ‘wholly foreign-owned’ and especially ‘wholly domestic-owned’ are close to zero. ‘Wholly owned’ has a higher frequency in the enTenTen corpus (16,383) than its alternatives; however, it ranks quite low among experts: Only 17% considered it completely appropriate, about 19% very appropriate, 25% moderately appropriate, 21% slightly appropriate, and 17% not at all appropriate. The two wordings that are considered very or completely appropriate by most experts, ‘entirely’ (75%) and ‘completely’ (73%), both have an extremely low frequency for ‘completely owned’ (287) and ‘entirely owned’ (201), according to the enTenTen corpus. ‘Totally owned’ has about the same frequency (214); however, it is very appropriate for only 31% of experts and moderately so for 31%. On the other hand, the

frequency of ‘all foreign-owned’ (927) is quite high given that it is not at all appropriate according to 50% of experts and only slightly appropriate for 28%.

The opinions of experts are divided in this case. Twelve out of 49 experts would choose ‘completely’, four of whom commented that it is the most commonly used wording (Code 4), and another noted that it is the most common in the context (Code 6). Moreover, one of the experts wrote that it is ‘simple, formal, clear’ (Code 2, Code 10, Code 8), and two other experts argued that it has the most relevant meaning – one of these experts specified that it ‘best describes that there are no other owners’ (Code 10).

Some experts also provided motivations for why other wordings are less appropriate. One expert wrote that ‘all’, ‘totally’ and ‘wholly’ do not fit very well with the word ‘partly’ in the second answer option (Code 5). Another expert said that ‘all’ ‘does not seem appropriate in this context’ (Code 5), while another commented that ‘it sounds bizarre and is not very catching’ (Code 9). Regarding ‘wholly’, three experts expressed concern that it would not be understood by everybody, one of whom even claimed that he ‘did not know this exists and what it means’ (Code 1). Moreover, it is a less commonly used wording according to two experts (Code 3) and ‘seems odd’ according to another (Code 9).

No disadvantages were listed for ‘entirely’ which would be the first choice of 10 experts. Although it is more difficult than ‘completely’, it is the ‘most precise’ (Code 8) and ‘fits better as a direct opposite to partly’ (Code 10). One of these experts also commented that the original wording, ‘wholly’, is ‘very unclear’ (Code 7) and ‘not common’ (Code 3).

However, there were also seven experts who put ‘wholly’ as their first choice. One explained that it is ‘as understandable as entirely’ (Code 2) but a ‘more correct term in this context’ (Code 10). Another expert claimed that it is ‘the most frequent collocate’ (Code 6); according to another, it is a ‘widely used term in the business environment’ (Code 6), but a disadvantage is that it requires respondents to have ‘knowledge of business affairs’ (Code 1).

Four experts would choose ‘totally’ – one commented that it is commonly used ‘in oral language’ (Code 4), and another that it is ‘commonly used in survey questions (answer modalities) so respondents are more used to this term’ (Code 6). There was also an

expert who would choose ‘all’ as it ‘suggests the whole quantity, group or thing and would be appropriate in the context of the question’ (Code 6 and Code 10), while ‘wholly would be used when referring to a person, thing or concept’ (Code 9).

In addition, 12 respondents could not decide among the options, and seven of them could not decide between ‘completely’ and ‘entirely’. One of them explained that they are both common words (Code 4), and another commented that the decision would depend on the intention of the question: ‘entirely in my view relates to the sum of parts’, while ‘completely refers more to the organisation as a whole, and relates to the fact that there may be multiple owners’. One expert also considered ‘totally’, along with ‘completely’ and ‘entirely’, as they ‘indicate an endpoint on a quantitative dimension’ (Code 10), while another would add ‘wholly’ instead because it is ‘used more often’ (Code 4); yet another would consider both of them but did not provide any motivation. One of the experts could not decide between ‘entirely’ and ‘wholly’, which ‘tends to be the choice of economists’, and another could not decide between ‘entirely’ and totally’.

Finally, two experts would not choose any wording but would completely reword the question. In particular, both problematised the use of the wordings ‘domestic’ and ‘foreign’, which are ‘very loaded descriptions’. They claimed that ‘domestic’ is ‘not a common expression in the UK’ and that in British English it means ‘a person who is employed to do housework’. One of the two experts also argued that ‘foreign’ is generally ‘a pejorative’.

In sum, it is difficult to make a final decision in this case. The original wording, ‘wholly owned’, has a higher corpora frequency and, as some experts explained, is a term commonly used in the context of business and economy. However, several experts pointed out that it is a wording that might not be familiar to all respondents. The wordings ‘entirely owned’ or ‘completely owned’ were better evaluated by experts, even though they have a much lower corpora frequency. Thus, the wording would benefit from further evaluations with different methods.

4.3.3.3 Laid down in the contract (W3)

‘Set in the contract’ is the wording most experts consider completely (37%) or very (29%) appropriate and has the second largest frequency according to the corpora (40). The only alternative with a higher frequency is ‘made in the contract’, although it is not

much higher (48); moreover, for experts, it is the worst alternative, as 62% of them considered it not at all appropriate, and an additional 25% considered it only slightly appropriate. The second choice, according to experts, would be 'established in the contract', which is completely appropriate for 8% of experts and very appropriate for 21%, while its wording frequency is only slightly lower (38) than for 'set'. The original wording in the Wageindicator questionnaire, 'laid down in the contract', is the second worst according to experts (only 6% consider it completely appropriate and 13% very appropriate) and the worst according to the corpus (16).

Twenty-five out of 49 experts put 'set in the contract' as their first choice, for which they provided several explanations. One said that it is 'the simplest and most common option' (Code 2 and Code 4), and another that it is often 'used in oral language' (Code 4). Two experts specified that the whole phrase is the most common – at least in American English, commented one – and the other wrote that it 'is the most normal way one would speak about this' (Code 6). Another expert noted that it seems 'more clear' (Code 8), and three experts made their choice based on what the term means, i.e., 'set is a more colloquial term', 'the hours in the contract are fixed' and 'indicated the number of hours which the contract specifies that you should work' (Code 10).

Some also explained why other alternatives are less appropriate. One said that they 'imply more of a negotiation', (Code 9), while 'set' implies that the contract is fixed. Moreover, the idiom 'laid down' is 'too complicated' (Code 1), 'seems cumbersome' (Code 9), and the phrase 'lay down hours' is very rarely (or even never) used (Code 5). According to two experts, 'made' is 'difficult to understand' (Code 1) as 'hours are not made' (Code 11), so it is grammatically incorrect. In addition, it refers 'more to the practice' (Code 9).

Six experts would choose 'established in contract', and one of them explained that this is because it 'included the idea of contract: the number of hours is established with the contract between the employer and the employee' (Code 10). Another expert noted that the original wording, 'laid down', is 'actually pretty funny in an unintended way – as it conjures up notions of laying down on the job' (Code 9). However, three experts put 'laid down' as their first choice, but they did not provide an explanation why. Moreover, one expert could not decide between 'set' and 'made', which are both 'common speech' (Code 4) and 'easier to understand' (Code 2).

Thirteen experts could not decide among the alternatives and suggested different wordings. Four of them would use the wording ‘specified in the contract’, and another four proposed that ‘set in’ should be replaced with ‘set out’ or ‘set by’. Moreover, two experts suggested using ‘stated in the contract’. There were several options proposed by only a single expert. For instance, one made the suggestion to ‘leave out the entire part’ so that the question would be ‘Do you usually work the number of hours in your contract?’ Another expert recommended rewording the question to ‘Do you usually work the number of hours you are contracted to?’ or using ‘agreed’ instead of ‘contracted’. Similarly, one expert advised using the form ‘agreed upon in your contract’, while another proposed replacing ‘laid down’ with ‘laid out’. Finally, the wordings ‘defined in the contract’, ‘fixed in the contract’, ‘described in the contract’ and ‘written in the contract’ were also suggested by experts.

To sum up, the phrase used in the master questionnaire, ‘laid down in the contract’, is not an appropriate wording, according to both the corpora and experts, but it is not clear what the ideal substitute is. Although ‘made in the contract’ has the highest frequency, it is not actually that high and, in particular, it is not that much higher than the other alternatives. Thus, more weight should be given to the opinions of experts who consider ‘made in the contract’ the worse alternative, as it is grammatically incorrect and difficult to understand. According to most experts, ‘set in the contract’ is the best option, as it is simple and has the most appropriate meaning; however, some argued that it should be replaced with ‘set out’ or ‘set by’. Moreover, several other alternatives were suggested which could be further analysed using text corpora and/or other methods.

4.3.3.4 Unforeseen problems (W4)

According to experts, the best wording is ‘unexpected problems’, which is considered completely appropriate by 55% and very appropriate by 45%; it is also the wording with the highest frequency in enTenTen (1,278). The original wording, ‘unforeseen problems’, comes second, according to both the corpora (1,004) and experts – 27% considered it completely appropriate and 39% very appropriate. The third place, according to the corpora and experts, goes to ‘unanticipated problems’ (21% completely appropriate and 38% very appropriate; enTenTen frequency: 328). No experts considered ‘out of the blue problems’ (7), ‘unlooked-for problems’ (0) and ‘unseen problems’ (100) completely appropriate. Although its frequency is quite high, the worst

wording according to experts is ‘unseen’, which 66% considered not at all appropriate and only 21% slightly appropriate. It is followed by ‘unlooked for’ (49% not at all appropriate and 40% slightly appropriate) and ‘out of the blue’ (50% not at all appropriate and 26% slightly appropriate).

Experts’ opinions were quite divided in this case. The most popular choice is ‘unexpected problems’, which was selected by 28 out of 49 experts. Five of them indicated that it is simple and easy to understand (Code 2), and two experts indicated that it is the most common (Code 4). In addition, two experts argued that it is univocal and has the right connotations, i.e., ‘a formal expression with no room for misinterpretation’ and ‘is clear and does not imply that solving them is unnecessary nor does it imply that people were lacking in planning’ (Code 8 and Code 10).

Some experts also explained why other wordings were less appropriate. First, ‘out of the blue’ is an expression that is not familiar to non-native speakers (Code 1) and is also too informal, according to one expert (Code 9). Second, the original wording, ‘unforeseen’, might not be understood by all people (Code 1), is ‘a little less common’ (Code 3), and is ‘less formal’ (Code 9). Third, ‘unlooked-for’ is ‘an awkward expression’ (Code 5) and ‘isn’t even English (sounds like an invention of a small child)’ (Code 3 and Code 11). Finally, ‘unseen’ ‘seems different than unexpected’ and ‘doesn’t convey the intended meaning’ (Code 9). No disadvantages were listed for the wording ‘unanticipated’.

On the other hand, six experts would choose ‘unforeseen problems’, and one of them commented that it ‘seems to win out slightly more’. Also, one of the experts would choose ‘unanticipated problems’ but did not explain why. However, most experts were undecided between different options. Four had difficulties deciding between ‘unanticipated’, ‘unexpected’ and ‘unforeseen’ – one of them explained that the options are ‘all fine’ and that ‘unanticipated is probably the most formal and unexpected the least formal’, so the decision depends on the required level of formality. Two of the experts could not decide between ‘unanticipated’ and ‘unexpected’, because the first ‘seems to be closest to what you’re trying to convey’ (Code 10) while the second is ‘a little simpler’ (Code 2). Finally, two could not decide between ‘unexpected’ and ‘unforeseen’ – one explained that ‘unforeseen reads better’ (Code 12) but ‘unexpected is more familiar to the respondents’ (Code 4).

To sum up, the original wording, ‘unforeseen problems’, is not the best choice according to the corpora and the experts. Some of them explained that it is less formal, less common, and might not be understood by everybody. According to the corpora and experts – although opinions were quite divided – the best choice is ‘unexpected’, which is the clearest, simplest and easiest to understand.

4.3.3.5 Sufficient energy/support (W5.1 and W5.2)

The preferred wording of experts is ‘enough’ (43% considered it completely appropriate and 38% very appropriate), which is also the best wording according to the enTenTen corpus: ‘enough energy’ (9,238) has a higher frequency than both the original ‘sufficient energy’ (1,645) and ‘adequate energy’ (577), while ‘enough support’ (3,894) has a higher frequency than ‘adequate support’ (2,383) and the original ‘sufficient support’ (1,336). ‘Sufficient’ is only the second choice (20% completely appropriate and 57% very appropriate), while ‘adequate’ is less appropriate (only 11% considered it completely appropriate and 33% very appropriate).

There were many different opinions regarding which is the most optimal choice and why. Twenty-three out of 49 experts would choose ‘enough’, eight of whom wrote that it is the most familiar and easiest to understand (Code 2). Three of them, as well as four others, also commented that it is the most common wording (Code 4), and one of them additionally noted that it is clear (Code 8). Moreover, another expert argued that it ‘has the most natural flow in both circumstances’. Four experts also explained why they found ‘adequate’ less appropriate. First, two experts argued that ‘adequate’ has a slightly different meaning (Code 9), and another specified that it ‘adds another dimension: respondents can say that they don’t have adequate energy because they have too much of it’ (Code 9). Finally, one expert argued that ‘adequate and sufficient both imply that I am not really getting enough, but I cope’ (Code 9).

Next, ‘sufficient’ was the first choice of 10 experts. One explained that ‘enough means it is more than sufficient’ and ‘adequate refers to a specific level (adequate level of energy)’ (Code 9), and thus it is ‘a bit weird’ in the context of this sentence. Similarly, another expert argued that although ‘enough’ is easier to understand than the other two, its meaning is different because it feels as if something is missing, i.e., ‘If I say “enough energy”, it could be the minimum of energy to do my job; if I say “enough support”,

one could say that there is never enough support' (Code 9). Another expert explained it differently: 'adequate and enough mean you have just enough energy to barely do your job' (Code 9), while 'sufficient implies that you have enough energy or support to do your job well or do your job in the way you would like to do it' (Code 10). Two experts commented that 'sufficient' is more formal (Code 10), and one added that it is a 'more positive word than adequate' (Code 10). Moreover, one of the experts noted that 'adequate' made him 'think about competence and not physical energy' (Code 9). On the other hand, two experts put 'adequate' as their first choice, but they did not provide any motivation except that 'it sounds better'.

In addition, one expert could not decide between 'adequate' and 'enough', and three could not decide between 'enough' and sufficient', but only two of them explained why, i.e., 'adequate has a different meaning' and is 'a bit ambiguous (it may be less than enough' (Code 7), while 'enough and sufficient are both fine'. Moreover, one expert was undecided between 'adequate' and 'sufficient' because they have different meanings, i.e., 'adequate denotes the average/typical level of energy or support', while the wording 'sufficient' 'addresses the lower threshold necessary to do the job.' Five experts had trouble deciding between all three options. Two of them explained that 'energy' and 'support' require a different selection. First, one expert argued that, for 'energy', any of the three would be fine; while for 'support', he suggested using 'adequate' if 'you are after a slightly wider assessment' as it 'implies support in both volume and quality', whereas 'enough' or 'sufficient' 'would more than likely be interpreted in terms of volume only'. Second, the other expert recommended rewording the first statement to 'I feel enthusiastic about my job' or the reverse 'I feel drained by my job', either of which would better convey the meaning than 'energy', which has connotations of 'electricity and other fuels' (Code 9). Regarding the second statement, he noted that the three options are not synonyms and the decision should be based on the appropriate level of strength: 'adequate means: just enough, barely enough' and 'enough is more positive and suggests that the level of support is quite good' or it 'might even suggest the supervisor is interfering a bit too much', while 'sufficient falls somewhere between the two'.

Finally, one expert would not select any option because 'none of these words makes a strong statement', explaining that to use the agreement scale, the 'statements need to be

unambiguously at either the positive or negative end of the conceptual continuum', and that if the respondents disagree, 'you don't know if he/she has more energy/support than needed or less.'

To sum up, the two wordings originally used in the Wageindicator, 'sufficient energy' and 'sufficient support', are less appropriate according to both the corpora and experts, although only one of the experts provided an explanation. The best wordings are 'enough energy' and 'enough support', which were better evaluated and are also more frequent in the text corpora. However, some experts warned that although 'enough' is the easiest, it is not appropriate because it has a different meaning. A final decision would likely require further research with other methods.

4.3.3.6 Sufficiently varied (W6)

'Varied enough' has the highest corpora frequency (419) and was considered completely appropriate by 27% of experts. The second best (according to the corpora), 'sufficiently varied' (55), was completely appropriate for 22% of experts, and 29% of experts considered it very appropriate, which is more than for 'varied enough'. The least appropriate choice, according to both experts and corpora, is 'adequately varied' – only 8% considered it completely appropriate and it has the lowest corpora frequency (1).

Again, opinions varied substantially and most experts were reluctant to choose only one option. Even though fewer experts considered it more completely appropriate than 'varied enough', some (19 out of 49) decided to put 'sufficiently varied' as their first choice. Four explained that it was because of its connotations, i.e., it is the 'least negative', 'a formal phrase', 'makes the most sense' and 'puts the emphasis on that concept' (Code 10), while one expert selected it 'purely on taste and how it sounds'. Four experts also explained why 'enough' is less appropriate: according to three of them, it is an incomplete, unfinished thought (Code 9), and according to the other, 'it isn't really grammatically correct' (Code 11). Regarding the appropriateness of 'adequately varied', only one expert noted that it 'really sounds very complicated' (Code 1), while another noted that 'it is not clear whether that is enough' (Code 9).

Only 15 out of 49 experts decided to put 'enough' as their first choice – three because it is the simplest (Code 2), four because it is the most common (Code 4), and one in order to be consistent with the previous question. Also, one of the experts commented that

‘enough sounds more like American English’. One of the experts indicated that ‘adequately’ has a different meaning, and one noted that ‘adequate’ is a property of a job (Code 9). Regarding ‘sufficiently varied’, none of these experts explained why it is less appropriate.

On the other hand, four experts put ‘adequately’ as their first choice – two explained that it was because ‘it sounds better’ or ‘reads better’, and another explained that an ‘adequately varied’ job implies ‘the variation needed for the job to be done’ (Code 10), while ‘sufficiently varied’ and ‘varied enough’ imply ‘that it could have been more varied’ (Code 9).

Three experts could not decide between ‘adequately’ and ‘sufficiently’, another two could not decide between ‘sufficiently’ and ‘enough’, and one could not decide between all three, but they did not provide any specific comments about why.

Finally, five experts would not select any of the offered choices. Same as for the previous case, one commented that the statement is not strong enough for an agreement scale. Another wrote that it was not clear if we meant tasks, work hours or contact with colleagues. Another expert criticised the assumption ‘that everyone would view a varied job as being a positive thing or something that they should value.’ Two experts suggested rewording the statement to ‘My job has enough variety’ or ‘My job has sufficient variety’.

In sum, as in the previous case, ‘varied enough’ seems to be a better wording than the original, ‘sufficiently varied’, as it has a higher text frequency and received a high evaluation by experts. However, more experts put ‘sufficiently varied’ as their first choice rather than ‘varied enough’. Some of them noted that using ‘enough’ is not grammatically correct in this sentence. To come to a final decision, the wordings would probably need to be evaluated with different methods.

4.3.3.7 Good state of repair (W7)

‘Good conditions’ has a much higher wording frequency (53,450) than ‘good state of repair’ (258), with 39% of experts considering it completely appropriate and 33% considering it very appropriate. On the other hand, ‘conditions’ are considered

completely appropriate by only 24% of respondents and very appropriate by 17% of them.

Thirty-four out of 49 experts would choose ‘conditions’ or its singular form ‘condition’ that was indicated by eight of them. Moreover, four argued that the plural form ‘conditions’ is not grammatically correct (Code 11). Three experts would pick ‘conditions’ because it is simpler and easier to understand (Code 2), three because it is often used (Code 4) – although ‘state of repair is probably more appropriate with machines and equipment’, noted one of them – and one because it ‘refers to a broader meaning construct’ (Code 10). Some also explained why they thought ‘state of repair’ is less appropriate. Seven indicated that it is too complicated, and five of them had never heard the expression before (Code 1). In addition, one expert argued that it ‘is ambiguous’ (Code 7), while another wrote that ‘it misleads me to think only about things that have previously been repaired’ (Code 9).

On the other hand, 11 experts would select the wording ‘state of repair’. One of them argued that it ‘is in common usage’ (Code 4), and another explained that it ‘suggests that the working order is good’ (Code 10), whereas ‘condition’ is a ‘broader concept and refers to the way the machine looks’ (Code 9). Moreover, four of them did not select ‘conditions’ because it is grammatically incorrect (Code 11); similarly, one noted that ‘the plural sounds non-native’.

Two experts could not decide between ‘conditions’ and ‘state of repair’. One explained that the two have different meanings: ‘good condition means that machines are good’, while ‘state of repair means that they have been bad, but are now repaired and good’. The other expert also stated that both words and the decision would depend on ‘the intended question meaning’. Moreover, he suggested that an alternative wording, such as ‘well maintained’, might also work. Finally, two experts would not select any wording. One of them just said that ‘it should be condition, not conditions’ but did not specify if it would be a better choice than ‘state of repair’. The other suggested using a different wording, either ‘in good working order’ or ‘are well maintained’.

To sum up, the original wording, ‘state of repair’, has a lower frequency and only a few experts considered it appropriate, as it is too complex and misleading. Instead, most

experts would choose the word ‘conditions’; however, some noted that it is not grammatically correct as it is and should be in its singular form: ‘condition’.

4.3.3.8 Staffing levels are sufficient (W8)

According to experts, ‘there is enough staff’ is the best wording: 31% considered it completely appropriate and 34% very appropriate. Its wording frequency in enTenTen (7) is only slightly lower than ‘staffing levels are adequate’ (8), which was completely appropriate for 20% of experts and very appropriate for 29%. The least preferable wording is ‘staffing levels are sufficient’, according to both the corpora (3) and experts (only 15% considered it completely appropriate and 31% very appropriate).

Twenty-two out of 49 experts would choose ‘enough’. Four of them commented that it was because it is the easiest to understand and ‘succinct’ (Code 2), three because it is more often used (Code 4), and another three because it is the clearest and ‘least likely to be misinterpreted’ (Code 8); one also wrote that it ‘has no negative connotations’ (Code 10). It should also be noted that one of the experts who would choose ‘enough’ commented that it should be corrected to the plural form: ‘there are enough staff’ (Code 11). On the other hand, ‘adequate’ and ‘sufficient’ ‘just seem odd’ (Code 9), according to one expert, and another commented, as in previous cases, that ‘adequate adds another dimension’ (Code 9). In addition, two experts did not problematise the use of ‘adequate’ or ‘sufficient’, but rather the phrase ‘staffing levels’, which is ‘a rather academic wording not familiar to many respondents’ and ‘seems too complicated’ (Code 1).

Ten experts would choose ‘staffing levels are adequate’, and three of them commented that it has the right connotations – one because we ‘refer to a level’, another because it ‘seems more effective related to staffing’, and another because it is ‘a better term for measuring quantity’. Moreover, another expert explained that ‘enough’ is less appropriate because it ‘indicates that it is almost more than enough’ and ‘sufficient’ because it ‘indicates it could have been more’ (Code 9). ‘Enough’ is also ‘a little less formal than the other two statements’ (Code 9), according to one expert. Eight experts would choose ‘staffing levels are sufficient’ – one of them explained that it ‘should be well understood’ (Code 2), and another one that ‘it is a more formal term’ (Code 10). In addition, one of them commented that ‘adequate’ is too ambiguous (Code 7) and ‘enough’ is ‘probably too colloquial’ (Code 9), while another also argued that ‘there is

enough staff is a broader concept' (Code 9). Moreover, two experts noted that it is grammatically incorrect – it should be worded 'there are enough staff' (Code 11).

On the other hand, four experts could not decide between 'adequate' and 'sufficient'. One explained, as in previous cases, that 'adequate denotes the average/typical level' and 'sufficient addresses the lower threshold', while another noted that if we were to use 'there is enough staff', we should also add something like 'to get all our work done'. Two experts were undecided between 'enough' and sufficient', where the decision should depend on what is intended, i.e., 'enough staff means everything is okay' and 'sufficient means we can work with this', while 'adequate is a property of a person not a rate' (Code 9). Moreover, one expert could not decide between all three options and noted that 'staffing levels are not the same as enough staff' as 'the latter only refers to quantity, while the first can refer to quality as well'.

Finally, two experts would not select any of the listed wordings. One repeated his comments from previous cases that the statement is not strong enough to allow the use of the agreement scale, as respondent disagreement can be interpreted in different ways. Similarly, the other expert also repeated his previous comments that 'adequate' means 'just enough, barely enough' and that 'sufficient' is a bit stronger and 'enough' is the strongest. The last, however, is not grammatically correct – it should be plural, i.e., 'there are enough staff' (Code 11).

In sum, the original wording 'staffing levels are sufficient' is again less appropriate than some of its alternatives. According to the corpora, the best wording is 'staffing levels are adequate'. However, its frequency is not very high and most experts prefer using 'there is enough staff'. This issue should probably be researched further with different methods.

4.3.4 Results for the PEW case

In this section, we perform the same analysis for the PEW questions as we did for the Wageindicator questionnaire. In total, 51 experts started responding but only 45 (88%) arrived at the last screen (demography). Twenty-five (56%) were non-native speakers, but 10 of them had lived in an English-speaking country for at least one year. The remaining 20 (44%) were native speakers: seven British, 12 American, and one Australian. The organisational affiliation for most of them was academic (55%), while

others worked in industry (23%) or elsewhere (e.g., government, public administration, statistical agency, non-profit organization). Many listed survey methodology or survey research as their main area of expertise, but some were more specific and listed questionnaire design or cognitive pre-testing of survey questions. In addition, there were also some who listed user experience, translation, statistics, linguistic quality control, quantitative research, nonresponse, mixed methods, labour economics or sociology as their main field.

Table 4.9 presents responses to the question, *'How appropriate is each wording if we want to make the corresponding question understandable to most people?'* (Question A in Figure 4.1). The best wordings according to expert reviews and the cells with over 25% of experts are shaded grey.

Table 4.9: Appropriateness of wordings

Wording	Completely appropriate	Very appropriate	Moderately appropriate	Slightly appropriate	Not at all appropriate	n
danger	3.9%	29.4%	37.3%	19.6%	9.8%	51
menace	0.0%	11.8%	21.6%	27.5%	39.2%	51
threat	60.8%	25.5%	5.9%	5.9%	2.0%	51
apprehensive	2.0%	5.9%	25.5%	27.5%	39.2%	51
concerned	37.3%	41.2%	11.8%	5.9%	3.9%	51
uneasy	0.0%	2.0%	25.5%	41.2%	31.4%	51
upset	0.0%	0.0%	9.8%	25.5%	64.7%	51
worried	39.2%	39.2%	15.7%	2.0%	3.9%	51
excused	2.0%	17.6%	33.3%	29.4%	17.6%	51
legitimate	6.0%	24.0%	36.0%	12.0%	22.0%	50
justified	47.1%	37.3%	11.8%	2.0%	2.0%	51
vindicated	0.0%	5.9%	17.6%	29.4%	47.1%	51
warranted	8.0%	24.0%	22.0%	24.0%	22.0%	50
chances	16.3%	24.5%	26.5%	18.4%	14.3%	49
probability	6.0%	22.0%	30.0%	32.0%	10.0%	50
risk	26.5%	44.9%	20.4%	6.1%	2.0%	49
are/be compassionate to	2.3%	2.3%	11.4%	29.5%	54.5%	44
are/be disposed to	0.0%	2.3%	20.9%	30.2%	46.5%	43
are/be in favour of	0.0%	13.6%	25.0%	27.3%	34.1%	44
are/be kind to	2.3%	0.0%	14.0%	18.6%	65.1%	43
support	45.5%	27.3%	18.2%	6.8%	2.3%	44
are/be supportive of	22.2%	40.0%	26.7%	4.4%	6.7%	45
are/be sympathetic to	6.7%	28.9%	26.7%	22.2%	15.6%	45
sympathize with	11.1%	24.4%	40.0%	17.8%	6.7%	45
abridging	0.0%	2.2%	10.9%	19.6%	67.4%	46

Wording	Completely appropriate	Very appropriate	Moderately appropriate	Slightly appropriate	Not at all appropriate	n
controlling	2.2%	13.0%	26.1%	13.0%	45.7%	46
curbing	4.4%	8.9%	17.8%	26.7%	42.2%	45
curtailing	8.7%	13.0%	28.3%	15.2%	34.8%	46
cutting back	9.3%	32.6%	18.6%	20.9%	18.6%	43
limiting	40.0%	37.8%	13.3%	6.7%	2.2%	45
restricting	35.6%	42.2%	20.0%	2.2%	0.0%	45
assembling	0.0%	6.8%	29.5%	29.5%	34.1%	44
collecting	52.3%	40.9%	6.8%	0.0%	0.0%	44
garnering	0.0%	2.3%	18.2%	15.9%	63.6%	44
gathering	31.8%	38.6%	22.7%	6.8%	0.0%	44
pulling together	4.8%	11.9%	21.4%	33.3%	28.6%	42
demanded money for the return of hostages	20.5%	25.0%	15.9%	11.4%	27.3%	44
demanded money for the return of sureties	0.0%	2.3%	13.6%	18.2%	65.9%	44
ransom money for hostages	45.5%	29.5%	18.2%	0.0%	6.8%	44
ransom money for sureties	0.0%	2.3%	15.9%	18.2%	63.6%	44
inclined	18.2%	34.1%	22.7%	15.9%	9.1%	44
prone	22.7%	34.1%	27.3%	6.8%	9.1%	44
advance	0.0%	2.3%	16.3%	32.6%	48.8%	43
boost	7.0%	9.3%	23.3%	20.9%	39.5%	43
encourage	40.9%	22.7%	25.0%	9.1%	2.3%	44
further	0.0%	4.7%	27.9%	34.9%	32.6%	43
promote	18.2%	54.5%	18.2%	4.5%	4.5%	44
afraid	9.5%	19.0%	21.4%	35.7%	14.3%	42
concerned	40.5%	40.5%	16.7%	2.4%	0.0%	42
preoccupied	2.4%	21.4%	19.0%	23.8%	33.3%	42
solicitous	0.0%	0.0%	12.5%	22.5%	65.0%	40
troubled	4.8%	19.0%	40.5%	19.0%	16.7%	42
worried	33.3%	33.3%	21.4%	11.9%	0.0%	42

Note: Shaded all cells over 20%

We interpret the results presented in Table 4.9 in the following subsections – each case has its own subsection (e.g., 'threat of terrorism' has Section 4.3.4.1). In addition, in each session, we analyse the responses to the open question where experts selected the wording they would choose and gave comments explaining why (Question B in Figure 4.1). Their comments were assigned odd (disadvantages) and even (advantages) codes in Table 4.8). Each comment could be assigned to more than one code.

4.3.4.1 Threat of terrorism (P1)

Most experts would use the wording ‘threat of terrorism’ (61% completely appropriate and 26% very appropriate), which is also the alternative with the highest frequency in enTenTen. ‘Danger of terrorism’ (116) and ‘menace of terrorism’ (126) have much lower frequencies and are considered less appropriate by experts.

Of the 49 experts who responded to this question, 39 would choose ‘threat of terrorism’, for which they listed various reasons. For instance, 11 wrote that ‘threat’ is a common word (Code 4), and seven explicitly mentioned that it is commonly used in the context of terrorism (Code 6). A few mentioned it also being common in mass media. Five experts said that ‘threat’ is the easiest for respondents to understand (Code 2), and one expert wrote that ‘threat’ is the clearest wording, which we interpreted and coded as meaning ‘univocal’ (Code 8). Then, we categorised eight responses as ‘right connotations’ (Code 10). Two experts said that ‘threat’ is ‘less loaded than others’, while the other six considered it as conveying the meaning intended by the question writers. Three of them provided general explanations, i.e., ‘it captures the overall intent of the question’, that it ‘gets the intended meaning across’ and is ‘closest to the conceptual ideal’; while the other three provided specific explanations, i.e., ‘it implies that terrorism is possible, but not necessarily existing yet’, that it ‘conveys the idea that it is an act that could have been carried out but has not, and also implies an intent on the part of the terrorists’, and as ‘the likelihood that a terrorist attack will happen’.

Some of the 49 experts also explained why they did not find ‘menace’ and ‘danger’ to be a less appropriate or not at all appropriate wording. Six discarded ‘menace’ because it is ‘a complicated concept’ and ‘sounds difficult for respondents with low education levels or who are not native English speakers’ (Code 1). Moreover, five experts wrote that it is not commonly used in language (Code 3), and one wrote that it is not common in the context of terrorism, which is something that two experts also said about ‘danger’ (Code 5). One expert wrote that ‘menace’ is ‘probably not understood in a consistent way’, and four said something similar for ‘danger’, i.e., it is ‘perhaps less semantically accurate’, and seems ‘a bit too general’ or ‘too vague’ (Code 7). Most reasons listed by experts, 11 for ‘menace’ and 14 for ‘danger’, were coded as ‘wrong connotations’ (Code 9). First of all, some wrote that both words are ‘too emotional, leading.’ or even ‘loaded’. Specifically, for ‘menace’, one expert wrote that ‘it applies to the emotional

fear that terrorism has on the general population' and 'danger' may have negative connotations – for instance, fear. On the other hand, others wrote differently about 'menace', i.e., that it 'is not strong enough', 'does not convey the same level of seriousness', 'makes the situation sound trivial', 'sounds as a nuisance', and implies 'something more along the lines of an annoyance or problem rather than something that could be a real danger'. Apart from the negative connotations for 'danger', they also wrote that 'it means something different in this context' and that it has a 'slightly different meaning' than the two other alternatives. More specifically, one of the experts wrote that it 'refers to latent insecurity' and another explained that it means 'trying to reduce the damages and not the threat of terrorism itself'. Similarly, another expert wrote that 'it sounds a bit like the question is asking about reducing how dangerous terrorism itself is, rather than reducing terrorism itself', and yet another that 'danger is getting a slightly different construct: the impact a terrorist attack would have if it were to happen, regardless of the likelihood that it happens at all'.

Among the remaining 10 experts, three would choose danger, 'as it seems to be the most familiar and easiest wording' and 'would probably be more accessible to a wider readership'. One expert could not decide between 'threat' and 'danger' because it would 'depend on the goals of measurement and what the question writer intended to measure'. One expert would choose 'menace' because 'danger is too vague, and threat has a negative connotation to it'. One expert suggested we should choose a different word – for instance, 'chance' – which 'is more neutral' compared to the three we offered, which are 'biased towards a negative reaction'.

The remaining four experts would not choose any wording and would rather completely reformulate the question. One suggested that we should ask: 'In general, how well do you think the US government is doing to reduce terrorism?' Another two said 'reducing terrorism', and one expert suggested that the question should also include the 'seriousness or intensity of terror'.

To sum up, according to both the text corpora frequencies and experts, the best wording is 'threat of terrorism'. It is very strongly endorsed by experts, as only a few did not put it as their first choice. However, some experts suggested that the question is itself biased and should be reworded, but this is beyond the scope of this chapter.

4.3.4.2 How worried (P2)

The experts would either keep the original wording, 'worried' (39% completely appropriate and 39% very appropriate), or choose 'concerned' (37% completely appropriate and 41% very appropriate). However, 'how concerned' has a slightly higher frequency (963) than 'how worried' (824). The only wording with a higher frequency is 'how upset' (1,696), for which 65% of experts said it was not at all appropriate, which is even worse than the two wordings with the lowest frequency, 'how apprehensive' (30) and 'how uneasy' (104).

Considering the responses to Question B, 20 out of 49 experts would choose the wording 'worried', six of whom because it is a 'basic word' that is the 'least formal option', 'more understandable to the general population' and 'can be understood by different groups of respondents (Code 2). Five experts wrote that it was because it is commonly and widely used (Code 4). Four listed reasons that we assigned to the 'right connotation' category (Code 10), i.e., is 'less loaded', 'does not change the meaning of the question too much' and 'has the emotional component we would expect in apprehension', which is presumably the target construct.

Some of the 20 experts also gave reasons why they would not choose the other offered alternatives. 'Concerned' is the word that was a second choice for many of the experts, as both 'worried' and 'concerned' are words 'that people use to describe these feelings' and 'should capture people's degree of apprehension'. However, it is too vague, explained three experts (Code 7). One of them pointed out that it is often used in these types of questions, but it 'triggers a lot of mistranslations in international surveys'. Moreover, three wrote that it is more loaded, 'restrictive in meaning' (Code 9) and 'too academic' (Code 1).

They also listed disadvantages for the other alternatives. Regarding 'apprehensive', five wrote that it is a difficult word to understand or that it is too formal (Code 1), while one expert commented that it is an 'uncommon term' (Code 3), and another said that it is 'less loaded' (Code 9). 'Uneasy' and 'upset' were ruled out because they are usually used to 'describe feelings for what has already happened' or 'that are a consequence of something' (Code 9), and also because they do not 'fit in the context' (Code 5). Moreover, 'upset' is too biased (Code 9) and vague (Code 7).

The second most appropriate wording is ‘concerned’, which would be the first choice of 14 experts. It is a widely used (Code 2) and common (Code 4) word that is ‘associated with this type of questioning’ and ‘the most consistent with the way people tend to talk about the issue’ (Code 6). Moreover, it is more objective, less leading, and expresses ‘the potential of something to happen’ (Code 10). One expert also said that it is the most grammatically appropriate word. ‘Apprehensive’ is too difficult (Code 2), while ‘uneasy’, ‘upset’ and ‘worried’ have wrong connotations (Code 9). In addition, ‘worried’ is also too vague (Code 7).

There were five respondents who could not decide between ‘concerned’ and ‘worried’, and some of them noted that the selection ‘would depend on the goals of measurement’. In contrast, there was also an expert whose first choice would be ‘apprehensive’, which ‘seems to strike the right chord here’ (Code 9), but also considered ‘concerned’ and ‘worried’ to be very appropriate. Finally, there was one expert who suggested rewording the question so that it does not use any of the suggested wordings: ‘In your mind, how likely is it that there will soon be another terrorist attack in the US?’

In sum, it is difficult to make a final decision in this case. On the one hand, wording frequencies suggest that ‘how upset’ is the best choice. However, most experts consider it not at all appropriate, as it is too vague and biased. Most experts would use either ‘how worried’, which was the first choice of two out of five experts, or ‘how concerned’, which was the first choice of two out of seven experts. Both wordings also have relatively high frequencies, and ‘concerned’ as a single word actually has the highest frequency, even higher than ‘upset’. On the other hand, some experts argue that it is too vague and biased and that it tends to be mistranslated in international surveys. Thus, further evaluation with different methods would be needed to make a final decision between ‘worried’ and ‘concerned’.

4.3.4.3 Ever justified (P3)

The original wording, ‘ever justified’, is the most appropriate wording according to both experts (47% completely appropriate and 37% very appropriate) and word frequency (227). Expert evaluations also correspond to corpora frequencies for the alternatives: ‘ever legitimate’ (33) and ‘ever warranted’ (31) are the second and third choice of

experts, respectively, while ‘ever excused’ (8) and ‘ever vindicated’ are the least appropriate wordings, respectively.

‘Justified’ was the first choice of 35 out of 49 experts. For 12 of them, one of the motivations was that it is the easiest to understand (Code 2). Moreover, four experts explained that it is a common word (Code 4), and one of them added ‘especially when speaking of philosophical topics’ (Code 6). Another three experts specified that it is ‘commonly used in the context’, ‘most consistent with the way most people talk about the issue’, or even ‘the only word that would typically be used in this context’ (Code 6). One expert wrote that it is ‘fairly clearly understood’ (Code 8) and explained that it ‘implies that it could be technically acceptable/legal’ (Code 10). However, one of the other experts provided a different explanation: ‘justification is about balancing the harms of torture versus the possible harms of a possible threat – acknowledging the evils of torture’ (Code 10).

The 35 experts also provided various explanations for why the other wordings are less appropriate or not appropriate at all. One expert stated that they would typically not be used in this context (Code 5), and another wrote that the ‘others are too emotional or leading’ (Code 9). Specifically, regarding ‘excused’, one expert noted that while ‘justified is about arguments’ (Code 10), ‘excused is about looking at the results’ (Code 9), which is probably not what we want to measure. Another wrote that although it is ‘widely understood and applicable’ (Code 2 and Code 4), ‘excusing would label the offender of torture of this crime as morally good’ (Code 9). Moreover, there were comments from seven other experts who also agreed that ‘excused’ ‘changes the context slightly’, i.e., ‘something could be unjustified but excusable’ and has wrong connotations, ‘requires the assumption that the use of torture is always bad, but that it might be deemed okay after the fact in specific circumstances’ or, in other words, it ‘implies that it should be technically wrong/illegal, but that we could ignore some violations’ or ‘that the action has already happened and it can be pardoned’ (Code 9). Regarding ‘legitimate’, four experts explained that the word is difficult or complex, especially for ‘respondents with lower education or non-native English speakers’ (Code 1). Moreover, it ‘implies an endorsement from some authority, which doesn’t seem appropriate for this question’ (Code 9). Two experts also noted that it is not grammatically correct, but ‘legitimised’ would be (Code 11). Next, ‘vindicated’ is also a

word that is ‘difficult to understand’ according to four experts, especially for non-native speakers and the lower educated (Code 1). Another expert also remarked that it is ‘used less frequently in everyday conversation’ (Code 3) and is ‘generally used post hoc (justified/cleared of suspicion after the event)’ (Code 5). In addition, three experts would not choose it because of its connotations, i.e., ‘suggests a further action that is taken to vindicate the action (i.e., some evidence of proof)’, ‘implies that it would be considered by others in some informal manner’ and is ‘often used in the connection with the clearing of a person or blame’ (Code 9). Finally, ‘warranted’ was considered difficult to understand by three experts (Code 1) and less widely used by three experts (Code 3). Another expert wrote that it is ‘clunky and does not flow in the question’ (Code 5), and another that it is ‘the worst of these, as it has multiple meanings’ (Code 7). Two other experts problematised the meaning as something that ‘could be warranted but not justified’ (Code 9).

However, there were three experts who put ‘warranted’ as their first choice, and two explained that it was because it ‘seems to be a more neutral word that doesn’t imply that the act is bad’ and ‘implies that torture can be necessary’ (Code 10). In addition, one expert could not decide between ‘legitimate’ and ‘warranted’, and another also put ‘legitimate’ in the mix, but did not provide any useful explanation. Moreover, six experts could not decide between ‘legitimate’ and justified’, but only one provided an explanation, i.e., ‘because both terms implicate that there needs to be good reasons for doing something’ (Code 10). Finally, two experts could not decide among the five options because ‘each word has a different meaning’ and ‘depends on what you want to know’.

In sum, most experts agree by far that ‘ever justified’ is the best wording, as already indicated by corpora frequencies. No objections were given against ‘justified’, so it is safe to trust this as the most appropriate wording.

4.3.4.4 Chances of attack (P4.1)

‘Risk of attack’ is the wording alternative with the higher frequency (483) and evaluation of appropriateness: 27% considered it completely appropriate and 45% very appropriate. The original wording, ‘chances of attack’, is the second best choice

according to both frequencies (49) and experts. ‘Probability of attack’ (28) was evaluated as the least appropriate.

Twenty-one out of 48 experts who responded to the question put ‘risk’ as their first choice. Five regarded ‘risk’ as being the easiest (Code 2), i.e., ‘a more basic English term’ which is ‘understandable to respondents’ and ‘easier for the general population to define’. Two of them added that they chose it because it is a more common word in the context (Code 6), one of whom specified ‘especially in combination with worry’ and the other that it is ‘more related to nuclear attack’. Both mentioned that an advantage is that it is ‘clear’ and ‘more concrete’ (Code 8). Moreover, six experts explained that it has the right connotations (Code 10), i.e., ‘more strongly implies that the action (nuclear attack) is negative’ or ‘something with a negative outcome’, and ‘captures both the probability aspect and the worrying aspect in one, thereby making the question a bit more about the concern of consequences’ or, as another expert put it, ‘risk is the danger of it happening’.

The 21 experts gave several reasons for excluding ‘chances’ or ‘probability’. Five experts considered ‘probability’ a ‘difficult concept’ that ‘may not be understood by all respondents’, ‘is too scientific’ and ‘people don’t think in probabilities’ (Code 1). One of them used the same motivation for not choosing ‘chances’. Moreover, another expert wrote that it ‘has too many different meanings (Code 7), and another that its singular form, ‘chance’, should be used (Code 11). In addition, wrong connotations were raised as a concern by four experts regarding ‘chances’ and five experts regarding ‘probability’. Both are ‘benign words, not suggesting a threat’ and you ‘cannot worry about probabilities or chances’ (Code 9). In fact, ‘worrying about the probability implies that you are concerned with the mathematical calculation of its chance’, i.e., ‘worrying about how to calculate the risk’. It also ‘suggests something that is quantifiable’. On the other hand, ‘chance’ implies low probability and ‘thus seems inconsistent with worrying about it’ (Code 9). Chance is also problematic in that it does not imply that the action is strongly negative enough because it ‘might have a connotation of a positive aspect’ (Code 9).

‘Chance’ would be the first choice of six experts because ‘chance is the term most often used in American English to represent probability’ (Code 6), and because it is ‘the simplest word’ that is ‘easily understood’ (Code 2) and ‘conveys the correct meaning’

(Code 10). They were motivated to not use ‘risk’ because it is less common in the context (Code 5) and ‘changes the meaning of the question a bit’ (Code 9). On the other hand, ‘probability’ would be the first choice of only three experts; even though ‘many people have a poor grasp of probability’ (Code 1), one of these experts specified that ‘probability is probably the most correct’ (Code 10) and ‘risk is not the same as probability’ as it ‘has a more negative connotation’ (Code 9).

In contrast, 11 experts suggested wordings that were not offered in the questionnaire. Four of them suggested using ‘possibility of attack’, two suggested ‘likelihood of attack’, and one suggested ‘potential for attack’. Moreover, they also criticised other aspects of the question: the vague quantifier ‘often’ can pose a difficulty for respondents, and the agreement scale is not appropriate – it would be better to measure frequency. For instance, ‘How often do you worry about possible nuclear attacks by terrorists?’, ‘How often do you worry about a nuclear attack by terrorists?’ or ‘How often do you worry there will be a nuclear attack by terrorists?’

There was one expert who wrote that all three are appropriate and could not decide between them, and there were three experts who could not decide between ‘risk’ and ‘chance’ but did not specify why. On the other hand, one expert could not decide between ‘risk’ and ‘probability’ because ‘although probability is a mathematical concept, it is widely known by the population’ (Code 4). Moreover, there was one expert who would not select any of the wordings because people ‘don’t worry about the probability, risk, whatever, but about the nuclear attacks themselves’.

To sum up, on the surface, ‘risk of attack’ seems the best choice as it is the wording with the highest corpora frequency and the one most favourably evaluated by experts. However, it should be noted that some experts argued that it changes the meaning of the question and is also not a common expression. Moreover, there are alternative wordings suggested by some experts that should be considered based on further evaluations using not only text frequencies and expert evaluations but other methods as well.

4.3.4.5 Sympathetic to terrorists (P4.2 and P4.3)

‘Support terrorists’ is the wording with both the highest enTenTen frequency (458) and the best evaluation by experts: 46% considered it completely appropriate and 27% very appropriate. The second choice is ‘be supportive of terrorists’, with a very low

frequency in the corpora (3). The original wording, ‘be sympathetic to terrorists’, has a slightly higher frequency (18) but a worse expert evaluation: 40% considered it only moderately appropriate, 24% very appropriate, and only 11% completely appropriate. The similar wording ‘sympathise with terrorists’ has only a slightly lower frequency (16), while ‘compassionate to terrorists’ (0), ‘disposed to terrorists’ (0), ‘in favour of terrorists’ (1), and ‘kind to terrorists’ (2) all have extremely low frequencies and are not considered appropriate by experts.

Twenty-two out of 46 experts who responded would choose ‘support terrorists’. Ten would select it because it is simple and the easiest to understand (Code 2). One of them warned that ‘support implies some type of active assistance’, but that it would depend on the intent of the question. Two of them added that it has a clear meaning, and one asserted that ‘support is the only word that most respondents would likely understand in a consistent way’ (Code 8). Nine experts explained their choice as having the correct connotation, i.e., ‘without emotion’ and being ‘the strongest word, ‘it includes a stronger level of support than the other options’, or that it ‘conveys the correct meaning’, i.e., ‘implies that you have provided money for a purpose or taken other actions that directly encourage terrorism’ or ‘that an individual may agree with a terrorist’s cause’ and because the question asks ‘about the limits to freedom of expression, therefore sympathise and support fit best’.

Various explanations were given for why other wordings are not appropriate. One expert said that all the other options seemed biased (Code 9). ‘Compassionate’ and ‘kind’ are ‘super odd’ (Code 9) and ‘may imply sharing (the) idea without however offering support’ (Code 9). Moreover, another expert said they both ‘convey a moral definition’, and another explained that they are completely inappropriate because ‘in theory it is possible to be against terrorism but compassionate/kind towards terrorists’; yet another explained that ‘you may feel compassion and kindness to someone or something, but not support their view’ (Code 9). The wording ‘disposed to terrorists’ is ‘too formal’ and ‘not often used in day-to-day conversation’ (Code 3); moreover, it ‘sounds awkward within the sentence’ (Code 5) and would usually mean ‘to get rid of something’ (Code 9). Then, ‘supportive to’ is less appropriate than ‘support’ because it does not make it ‘clear if someone provided resources or not to a person or cause’ (Code 7) and, being longer, it ‘will increase the reading level which may make the question

more difficult for some respondents (Code 13) – the same also holds true for 'sympathetic to'. Finally, 'sympathetic to' and 'sympathise with' are less clear than 'support' (Code 7) and have different implications, i.e., 'sympathise means that you understand and feel for the person or cause' (Code 9).

Six experts would choose 'supportive of', of which two explained that it is 'a bit clearer than disposed towards' (Code 8) and 'is more grammatically correct and less emotional' than 'support' (Code 10). Only one expert put 'sympathetic to' as their first choice, explaining that it is less strong than 'support', which 'implies giving aid or money' (Code 9). Moreover, four experts would choose 'sympathise with' but only one provided an explanation, i.e., because it expresses 'the fact that somebody thinks positively about something/somebody without the necessity of being actively supportive' (Code 10). The same is true for 'supportive of'.

Seven experts could not decide between two or three options. For instance, two could not decide between 'support' and 'supportive of', both of which are 'more precise than the other options' (Code 8), which are too vague. Two experts could not decide between 'support' and 'sympathise' but did not provide any comments about why. Another expert could not decide between 'supportive of' and 'sympathetic' and offered an alternative choice: 'well-disposed to terrorists'. Finally, one of the experts could not decide between 'sympathetic' and both forms of the 'support' wording – the decision would depend on the intended meaning, i.e., 'if they are just of a kind disposition towards those people', then 'sympathetic to', and if it is 'about people who don't just have sympathies to those groups', then 'supportive of' or 'support' is the right choice.

Six experts could not decide among any of the options. Two of them had difficulties because the decision would have to depend on the intent of the question. One specified that with 'support' there is the 'implication that you're asking about groups that have a hands-on role in supporting terrorist groups, whether it be with finances of weapons or advice or whatever', while other acceptable alternatives 'ask about different levels of involvement and seem to be more distant'. Moreover, another expert noted that the question 'has a stranger word order' as 'without court order' could be read as 'belonging to the terrorists, not the police searching houses'. One expert suggested using 'sympathise and support terrorists' and argued that although it makes the sentence longer 'this is such a delicate issue concerning very basic human rights, that it needs it'.

Finally, two of them would not choose any of the offered alternatives and would instead form the question completely differently: one criticised the agree–disagree format of the question, and the other noted that it 'seems to have a hidden proposal that terrorism is bad and that, by extension, everybody who is not against is in favour', concluding that the question 'would have needed a lot of discussion if put in a TRAPD process', which is probably the acronym for translation, review, adjudication, pre-testing and documentation.

To sum up, the results indicate that the best wording is 'support', which has the highest frequency and would be the first choice of half of the experts. However, other experts argued that it is too strong and emotional, and has the wrong connotations. In addition, some remarked that the decision should depend on what the question designers intended. In fact, the original wording used in the PEW questionnaire is 'sympathetic to', which is not exactly a synonym of 'support'. Thus, this case could also benefit from some further evaluations using other methods.

4.3.4.6 Restricting liberties (P4.4)

'Limiting liberties' has a low frequency (4) but is considered completely appropriate (40%) or very appropriate (38%) by most experts. The original wording, 'restricting liberties', has only a slightly higher frequency (7) and is the second best evaluated: 36% said it is completely appropriate and 42% very appropriate. 'Abridging liberties' (1), 'controlling liberties' (0), 'curbing liberties' (1), 'curtailing liberties' (6) and 'cutting back liberties' all have lower enTenTen frequencies and are considered not at all appropriate by most experts.

Experts' opinions varied a lot for this case. However, 16 out of 44 agreed that 'restricting' is the best choice. Five experts believed it is simple and widely understood (Code 2), and two believed it is a common word (Code 4), one of whom also stressed that it 'is the only option that sounds like a natural wording in this context' (Code 6). Similarly, another expert wrote that 'civil liberties is a concept that is commonly discussed (e.g., on television)', and another wrote that its use 'is consistent with the way this issue is typically discussed in society' (Code 6). In addition, one expert argued that it 'is the most accurate and clearest term' (Code 8).

Several reasons were provided for why other alternatives are less appropriate – except for ‘limiting’, for which no disadvantages were listed. One expert wrote that except for ‘restricting’ and ‘limiting’, most alternatives are too complex (Code 1); while another limited the selection to ‘curbing’, ‘curtailing’, and ‘abridging’ as words that ‘would not be understood by all’ (Code 1). Another expert regarded ‘curtailing’ as being less widely used (Code 1) and understood (Code 3). Yet another expert considered ‘curtailing’, ‘abridging’ and ‘curbing’ as ‘too uncommon’ (Code 3). In addition, some experts found flaws with the connotations, i.e., ‘curbing is too informal’ and ‘curtailing is almost too formal for this question’ (Code 9), and also meanings, i.e., ‘controlling’, which ‘suggests that people still have liberties but the government controls them, which doesn’t make sense’ (Code 9).

For seven experts, ‘limiting’ was the preferred and only choice. Three explained that it is the easiest to understand (Code 1), and one of them added that its meaning is ‘quite clear’ (Code 8); another expert reasoned that ‘limiting is used most often in media commentary on civil liberties’ (Code 6). One of these experts wrote that the other alternatives are ‘too biased’ (Code 9). Three experts put ‘curtailing’ as their first choice. One did not provide a motivation, one commented that it seems to be ‘the right collocation’ (Code 6), and one wrote that along with ‘limiting’ and ‘restricting’, it is the most appropriate choice because of its meaning, i.e., it ‘suggests a reduction in something that currently exists’ (Code 10).

Fifteen experts could not decide among two or three of the alternatives, eight of whom were undecided between ‘limiting’ and ‘restricting’. Two of them explained that the two are common, everyday words (Code 4) that imply the correct meaning (Code 10), and another expert wrote that they are ‘pretty clear’ (Code 8). Other wordings might be unfamiliar to some non-native speakers, one expert wrote, or even average Americans, another noted (Code 1). On the other hand, two of the experts also considered ‘cutting back’, along with ‘limiting’ and ‘restricting’, but did not provide an explanation for why, except that ‘other words are rather difficult’ (Code 1). Moreover, one of the experts considered ‘controlling’ appropriate, along with ‘limiting and ‘restricting’, while ‘abridging’ and ‘curbing’ were not known to him as a non-native speaker (Code 1). Yet another expert considered ‘curbing’ alongside ‘limiting’ and ‘restricting’, as they all ‘suggest something other than the complete ending of the average person’s civil

liberties, which is more likely to be the interpretation if you use curtailing' (Code 10 and Code 9). One expert could not decide between 'limiting' and 'cutting back', both of which are 'the closest to the intention of the question' (Code 10), and another could not decide between 'limiting' and 'curbing', but their provided explanations are not very clear. Finally, one of the experts suggested two alternatives that were not offered in the questionnaire: 'reducing civil liberties' and 'harming civil liberties'.

To sum up, 'restricting liberties' has the highest corpora frequency and is the first choice of four out of 11 experts. However, 'limiting' was slightly better evaluated (even though fewer than two of 11 experts would put it as a first choice). No disadvantages were listed for either of the two. Although both options would likely be fine in this context, further evaluations with other methods would be beneficial. In particular, other alternative wordings suggested by experts should be evaluated.

4.3.4.7 Collecting information (P4.5)

The original wording, 'collecting information', has only the second highest wording frequency (4,417) but the best evaluation by experts: 52% considered it completely appropriate and 41% very appropriate. The wording with the higher frequency, 'gathering information' (7,542), came second according to experts (32% said it was completely appropriate and 39% very appropriate). The third choice is 'assembling information', according to both experts and frequencies (83), while the fourth is 'pulling together information' (48) and the worst, 'garnering information' (37).

Twenty-six out of 43 experts would choose 'collecting'. Ten would choose this word because it is the simplest and easiest to understand (Code 2); three wrote that it is a common term (Code 4), while another three specified that it is common in the context of data and information, i.e., 'standard term' or 'the most common usage in US media for this topic' (Code 6). For eight experts, it also has the clearest, least confusing meaning (Code 8), while the others seem 'too vague'. Four experts argued that it is 'the word that conveys the meaning properly', 'without losing formality', and that it implies 'a bit more intentionality than simply gathering information' and 'a real action of the government' (Code 10).

Various motivations were given for not choosing other alternatives. One expert said that all of the others 'seem too vague' (Code 7). Moreover, 'assembling' and 'garnering' are

‘a little too complex’ and ‘don’t work for a general audience’ – one expert did not even know the words (Code 1). Also, they are ‘probably less frequently used’ (Code 3), in particular in the context of information (Code 5). ‘Pulling together’ also would not work for a general audience (Code 1) and in this context (Code 5) because it ‘refers to real people’ (Code 9). For two experts, it is also ‘too informal’ or ‘unnecessarily basic’ (Code 9). ‘Garnering’, on the other hand, is a simple and common word that also seems to be appropriate, but one expert argued that it is ‘used less frequently in everyday conversation’ (Code 3), and another that it ‘can be understood to mean that it is personal (like a private investigator) rather than a routine operation applied to all citizens’ (Code 9).

‘Gathering’ was the second most popular – the first choice of five experts. One of them explained that he regularly heard about ‘intelligence-gathering’ (Code 6), ‘which is a closely-related concept’ (Code 10). However, these experts did not reveal any motivations for not choosing any of the other wordings. There was one expert who would choose ‘pulling together’ because it ‘is the closest in meaning to assembling’ (Code 10) ‘without using a difficult word’ (Code 2). In addition, one of the experts would choose ‘assembling’ but did not provide an explanation why.

Finally, 10 experts could not decide between ‘collecting’ and ‘gathering’, as they are ‘the only simple enough ones to use for the whole population’ (Code 2), very common words (Code 4), in particular in the context of information (Code 6), ‘semantically and pragmatically best fit the question’, and imply ‘that there is no performed checklist that you are trying to get all the pieces of’ (Code 10).

To sum up, it is not clear if the best alternative is ‘gathering information’, which has the higher frequency, or ‘collecting information’, which is preferred by most experts. No disadvantages were given for either of the two, except for one expert who argued that ‘there is more intentionality’ in ‘collecting’, while ‘gathering’ is more passive. In any case, these two alternatives should be further researched with other methods.

4.3.4.8 Ransom money for hostages (P5)

According to the corpora, the best combination is to use ‘ransom for hostages’, which has both the highest frequency in enTenTen and is the best according to experts (46% completely appropriate and 30% very appropriate); while the original wording, ‘ransom

money for hostages', is longer and has a frequency of only one in the corpus. Both are better than the alternative 'demanded money for hostages', which has zero frequency in the text corpora. However, almost half of the experts considered it appropriate (21% completely appropriate and 25% very appropriate); in any case, 'hostages' is a better wording than 'sureties', which does not collocate with ransom and which the experts considered not at all appropriate.

Twenty-nine out of 45 experts would choose 'ransom money for hostages'. Five would choose them because both 'ransom' and 'hostage' are well understood (Code 2), while two experts wrote that they are both common words (Code 4). Moreover, seven experts commented that the whole context, 'ransom money for hostages', is very common, i.e., 'consistent with typical speech' and 'used a lot on TV, movies and news' (Code 6).

Experts provided many reasons for not using 'demanded money' and 'sureties'. First of all, although being 'probably more specifically correct in terms of terminology' (Code 10), 'demanded money' is less commonly used in this context, argued one expert, and six others also commented that it is uncommon. For instance, one expert wrote that he 'hardly ever used that term', and another that he has 'never heard of the phrase 'demanded money'; yet another commented that it is 'not a term common outside the playground and soap opera' (Code 5). Two experts noted that 'demanded money' is 'unnecessarily complex' and makes the 'question long and therefore more difficult to comprehend' (Code 1), while another expert commented that it 'may confuse some people' (Code 7). Finally, two of the experts argued that it is grammatically incorrect (Code 11). Even more disliked is the wording 'sureties', which seven experts considered too difficult and five had actually 'never heard the term sureties' and did not know what it meant (Code 1). Also, others commented that it is 'too technical', 'not typically used in most participants' vocabulary' and 'many people would not understand what it meant' (Code 1). Moreover, six commented that it is not common – one of them specified that it is 'not commonly used in American English', and one even wrote that it is 'a bit obscure' (Code 3). For one expert, it was not an uncommon word but it was uncommon in the context (Code 5), arguing that 'surety is a debt obligation' and that 'to pay money for a surety is to seek to obtain a document and an obligation, not a hostage' (Code 9).

On the other hand, there were 10 experts who would choose the wording ‘demanded money’, which is ‘understandable to anyone’ (Code 2), ‘commonly used’ (Code 4) and ‘commonly heard in the media’ (Code 6). In addition, it is also ‘more concrete’ and more ‘straightforward’ (Code 8). ‘Ransom’ was not known to one of the experts, and two others also commented that it is more complex, less understandable, and ‘might not be clear to all respondents’ (Code 1).

There were four experts who could not decide between ‘ransom’ and ‘demanded’ but did not provide any specific comments. Finally, one expert suggested an alternative wording that was not suggested in the questionnaire: ‘it (the government) NEVER pays money to terrorist groups in exchange for setting hostages free’.

In summary, the original wording, ‘ransom money for hostages’, is the best alternative according to both the text corpora and experts. However, three experts argued that ‘ransom’ might not be understood by all respondents. Their assumption should be tested using other methods. Regarding ‘hostages’, on the other hand, there is no doubt that it is the only appropriate choice, as several experts have claimed that the wording ‘sureties’ is unknown to them.

4.3.4.9 Prone to violence (P6)

The original wording, ‘prone to violence’, is the most frequent (452) and best evaluated by experts: 23% considered it completely appropriate and 34% very appropriate. The alternative wording, ‘inclined to violence’, has a lower frequency (39) and is considered a bit less appropriate by experts (only 18% considered it completely appropriate and 34% very appropriate).

Sixteen out of 43 experts would choose the wording ‘prone to violence’, which ‘is a common type of expression in US English’ (Code 6) according to one of the experts, ‘makes the most sense’ (Code 10) according to another, and is a ‘turn of phrase’ according to a third, which means that it is an idiomatic way of saying it. On the other hand, they did not provide almost any argument for why ‘inclined’ is less appropriate. Only one of the experts noted that, in this case, the stimulus is less strong (Code 9).

There were actually 11 experts that would choose the wording ‘inclined to violence’, with four arguing that it has the correct implications, i.e., ‘implies something accidental

rather than deliberate' (Code 10). Moreover, they specifically compared its meaning and implications with the wording 'prone'. First, one expert wrote that 'prone implies that they act more violently' (Code 9), while 'inclined means they may act but also could support a belief or cause' (Code 10). Second, one expert wrote that it 'implies that such religions are likely to sanction violence' (Code 10), 'whereas prone to violence seems to suggest that violence is more likely to happen within this religion' (Code 9). Third, and similarly, one expert commented that 'prone is more extreme than inclined' (Code 9) and explained that since the second sentence ('all religions are about the same when it comes to violence') is less extreme because of the term 'about', 'inclined' should therefore be used so that the level of extremity would be the same in both statements (Code 10).

On the other hand, seven experts could not choose between 'inclined' and prone' but did not provide any specific motivations. One could not decide between the two and suggested an alternative wording: 'Some religions tend to be more violent than others'. Finally, two experts, although considering 'prone' more appropriate than 'inclined' because it is more common (Code 4) and less complex (Code 1), also suggested their own alternatives: one would word it 'some religions are more likely to be violent than others', and the other 'some religions are known for their violence more than others'.

To sum up, on the surface, it seems there is not much to discuss here, as the original wording, 'prone to violence', has a higher enTenTen frequency and is better evaluated by experts than the alternative, 'inclined to violence'. However, there are also quite a few experts who would choose 'inclined' or were undecided between the two options, for which they provided good motivations. In particular, 'prone' is stronger and more extreme. Thus, it might be beneficial to evaluate the two wordings alongside some other methods.

4.3.4.10 Encourage violence (P7)

'Promote violence' is the wording with the highest frequency (1,002), followed by 'further violence' (922). However, 'further' is not at all appropriate (33%) or only slightly appropriate (35%) according to experts, and 'promote' was completely appropriate for only 18% and very appropriate for 54% of experts. The original wording, 'encourage violence', was considered completely appropriate by 41% of

experts and very appropriate by 23%; however, the wording frequency is notably lower (517) than the first two. ‘Advance violence’ (4) and ‘boost violence’ (2) both have low frequencies and low expert evaluations.

Fourteen out of 42 experts would choose ‘promote’, two because it is clear (Code 8), simple and easy to understand (Code 2), another two because it is a common word (Code 4), and yet another two because it is the most commonly used in the context of violence, and one specifically mentioned in the US media (Code 6). Moreover, four experts argued that it ‘better captures the aim of the question’, that it ‘implies positive action to support a belief or cause’, or that it ‘implies that increases in violence do not have to be explicitly forwarded by religious leaders – it includes the possibility that this is a by-product of the religion rather than an explicit part of its teachings’ (Code 10).

Various reasons were provided for why the other alternatives are less appropriate. One expert indicated that except for ‘encourage’ and ‘promote’, all the other words ‘aren’t very specific and are overly basic’ (Code 7). In addition, two experts commented that ‘boost seems to be more colloquial’ or it ‘sounds like popular language’ (Code 9). Next, ‘encourage’ is less appropriate because it ‘doesn’t apply action, simply thoughts that concur with a belief or cause’. Moreover, ‘encourage’ is usually used in relation to somebody doing or being something (Code 11).

However, 10 of the experts would choose ‘encourage’ because it is ‘easily understood’ (Code 2), ‘is a much more common word to use’ (Code 4), it ‘collocates best with violence’ (Code 6) and ‘is more neutral’ (Code 10). What makes ‘promote’ less appropriate is that it tends ‘to be positive’, according to one of them, while it is ‘too definitive’ according to another; yet another argued that ‘it tends to convey the notion of promotion as in advertising’ (Code 10). Two of the experts would choose ‘boost’ – one believed that ‘boost is understandable, while the other added that, in order to make it clearer, the sentence should be reworded to ‘The Islamic religion does not boost violence more than other religions do’.

On the other hand, 10 other experts could not decide between ‘encourage’ and ‘promote’, finding them to be equally appropriate and valid. One explained that they ‘are the most widely used’ (Code 4) and ‘used in the context of persuading and influencing’ (Code 6). There was one expert who could not decide between ‘boost’ and

‘promote’, and another who could not decide between ‘boost’ and ‘encourage’, but they did not comment on this indecision. Finally, there were two experts who did not choose any wording. One warned that we should ‘be very careful with such a question, out of context in particular’ and another labelled it as an ‘Islamophobic question’. He suggested asking something like ‘Please look at the religions/beliefs listed below and tell me the extent to which each promotes violence amongst its believers: Strongly Promotes, Promotes, Neither Promotes nor Discourages, Discourages, Strongly Discourages’.

In summary, the original wording, ‘encourage violence’, is the one best evaluated by experts, and one-third put it as their first choice. On the other hand, ‘promote violence’ and ‘further violence’ have higher wording frequencies. Moreover, some experts put ‘promote’ as their first choice, and some could not decide between ‘promote’ and ‘encourage’. Several motivations were given for both, but it is hard to make a decision at this point. The case should be further evaluated by other methods.

4.3.4.11 Concerned about extremism (P8)

The original wording, ‘concerned about extremism’, is the only wording with a frequency higher than zero (3) and is also the favourite wording for experts: 41% considered it completely appropriate and 41% very appropriate. The second choice of experts is ‘worried about extremism’ (0), which was completely appropriate for 33% and very appropriate for 33%. ‘Troubled about extremism’ and ‘afraid about extremism’ are only moderately or slightly appropriate according to experts, while ‘preoccupied about extremism’ and ‘solicitous about extremism’ are not at all appropriate for most of the experts.

Fourteen out of 36 experts who responded to this question would choose ‘concern’. First, one commented that it ‘is the natural/obvious choice’ as it is well understood (Code 2). Second, one expert argued that it is the word to be used in the context ‘about societal developments in general’, and another that it ‘is the most consistent with the way people typically talk about this issue’ (Code 6). Third, six experts explained that it has the right connotations. Specifically, it is a ‘less strong’ term, ‘probably more neutral’, ‘seems less judgmental’ and ‘carries connotations of both thinking and feeling, which both seem in scope’. Moreover, as in one of the earlier cases where we had

‘concerned’, one expert explained that it ‘means you’re thinking about it, but not necessarily negatively’.

Some experts also explained why other alternatives are less appropriate. One explained that you are ‘afraid’ or ‘worried’ about something ‘if you are personally involved’, which is not the case in the given example since it is about societal developments (Code 5). In addition, ‘afraid’ changes the meaning of the question – it ‘sounds like something will definitely happen’ (Code 9). Then, ‘solicitous’ also has a wrong connotation, as it is usually ‘used in terms of caring for someone’ (Code 9). Next, ‘preoccupied’ and ‘troubled’ are too strong as they sound ‘like someone is constantly thinking about something’ (Code 9), argued one of the experts, while another said nearly the same thing about ‘preoccupied’. Also, the meaning of ‘worried’ and ‘troubled’ is not exactly equivalent, i.e., ‘I could be concerned without being worried if I was just the kind of person who does not worry, no matter how much of a concern something is’ (Code 9).

Four experts put ‘worried’ as their first choice, and two of them provided an explanation – one wrote that it is because it implies ‘negative thoughts, but not so much that they are debilitating to your daily life.’ Two experts picked ‘preoccupied’ – one of whom because ‘it is an understandable word to use in this item’, and the other because it is ‘more extreme’ and as such is needed in such a statement. Moreover, one expert would choose ‘afraid’ but did not explain why, and another would choose ‘troubled’, which is ‘more basic’ (Code 2). Finally, eight experts could not decide between two or more of the listed wordings, and one expert suggested rewording the question so that it would be more neutral.

To sum up, the decision seems pretty straightforward here: The original wording, ‘concerned about extremism’, has the highest corpora frequency and is the first choice of most experts. No objections were given by any experts regarding this combination. Thus, we recommend keeping the older wording.

4.3.5 General criticism of the methodology

In addition to comments to specific questions, several experts used the comment box to provide comments that did not directly address the question we asked but provided other suggestions on how to improve the question. Most of the comments related to

wording and format changes, which we already evaluated in the previous subsection for each case (survey question) separately.

However, a few of these comments were addressed to shortcomings in the methodology we used. First, some pointed out that it is not clear what we mean by appropriateness. Second, some experts pointed out that the wordings we listed were not synonymous and the decision would depend on what concept we want to measure. Third, many experts pointed out that the proposed wording variations will not have any effect on question quality – other aspects of question wording should be addressed so that we could observe any relevant changes. Finally, a few respondents who were non-native English speakers did not consider themselves to be good evaluators – we will look into this issue in the next subsection, where we compare the evaluations made by native and non-native speakers.

4.3.6 Comparison of native and non-native speakers

Thirty-one out of the 51 experts (61%) who participated in the Wageindicator questionnaire evaluation are non-native speakers, and 24 out of the 49 experts (49%) who responded to the PEW questionnaire are non-native speakers. In this subsection, we compare their responses to Question A (see Figure 4.1), which we recoded to three categories by merging 'not at all appropriate' and 'slightly appropriate' (1 and 2) and 'very appropriate' and 'completely appropriate' (4 and 5).

The sample is very small and not representative, so the analysis will only be descriptive and will not compute any statistical differences or other measures.

4.3.6.1 The Wageindicator case

In Table 4.10, we compare the percentages of responses between native and non-native speakers for the Wageindicator questionnaire.

Table 4.10: Comparison of non-native and native speakers in the Wageindicator evaluation

	Non-native (n=31)			Native (n=18)		
	Very and Completely	Moderately	Not at all and Slightly	Very and Completely	Moderately	Not at all and Slightly
form	23%	35%	42%	0%	44%	56%

	Non-native (n=31)			Native (n=18)		
	Very and Completely	Moderately	Not at all and Slightly	Very and Completely	Moderately	Not at all and Slightly
kind	83%	13%	3%	67%	17%	17%
type	83%	17%	0%	100%	0%	0%
variety	3%	10%	87%	0%	17%	83%
all	7%	14%	79%	6%	17%	78%
completely	70%	23%	7%	78%	6%	17%
entirely	77%	20%	3%	72%	17%	11%
totally	53%	30%	17%	39%	33%	28%
wholly	31%	24%	45%	44%	28%	28%
established	41%	28%	31%	11%	33%	56%
laid down	28%	45%	28%	6%	11%	83%
made	3%	17%	79%	0%	0%	100%
set	65%	26%	10%	67%	28%	6%
out of the blue	4%	25%	71%	0%	17%	83%
unanticipated	57%	40%	3%	61%	28%	11%
unexpected	100%	0%	0%	100%	0%	0%
unforeseen	68%	29%	3%	61%	33%	6%
unlooked-for	3%	10%	86%	0%	6%	94%
unseen	10%	3%	86%	0%	11%	89%
adequate	41%	24%	34%	44%	33%	17%
enough	83%	10%	7%	72%	17%	6%
sufficient	79%	21%	0%	67%	17%	11%
adequately varied	27%	40%	33%	17%	28%	56%
sufficiently varied	52%	35%	13%	50%	28%	22%
varied enough	57%	20%	23%	22%	28%	50%
conditions	84%	6%	10%	50%	6%	44%
state of repair	29%	36%	36%	61%	28%	11%
Staffing levels are adequate.	48%	29%	23%	50%	28%	22%
Staffing levels are sufficient.	37%	50%	13%	61%	33%	6%
There is enough staff.	73%	20%	7%	56%	22%	22%

In most cases, we see that there are important differences between expert evaluations by native and non-native speakers. First, it can be observed that non-native speakers tend to prefer the wordings ‘kind of contract’ and ‘form of contract’ more than native speakers, while they endorsed ‘type of contract’ less strongly than native speakers (W1). Second, native speakers tend to evaluate ‘wholly’ more favourably than non-native speakers (W2). Third, the wordings ‘established’, ‘laid down,’ and ‘made in the contract’ are better evaluated by non-native speakers, while ‘set in the contract’ is the wording that more native speakers prefer (W3). Fourth, native speakers tend to appreciate the

wording ‘unanticipated’ more than non-native speakers, while the latter seem to better evaluate the wording ‘unseen’ (W4). Fifth, more non-native speakers consider ‘enough’ and ‘sufficient’ very or completely appropriate than natives, while the latter give a slightly better rating to ‘adequate’ (W5.1 and W5.2). Sixth, and similarly, ‘varied enough’ is more popular among non-native speakers than native speakers (W6). Seventh, the strongest difference between native and non-native speakers is in regards to the wording ‘working conditions’, which is seen as more appropriate by non-native speakers, while native speakers prefer ‘state of repair’ (W7). Eighth, native speakers show a strong preference for the wording ‘staffing levels are sufficient’, while non-native speakers prefer the wording ‘there is enough staff’ (W8).

4.3.6.2 The PEW case

Table 4.11, on the other hand, shows the comparison of non-native and native speakers for the PEW questionnaire.

Table 4.11: Comparison of non-native and native speakers in the PEW evaluation

	Non-native (n=24)			Native (n=20)		
	Very and Completely	Moderately	Not at all and Slightly	Very and Completely	Moderately	Not at all and Slightly
danger	38%	38%	25%	30%	40%	30%
menace	17%	29%	54%	0%	15%	85%
threat	79%	8%	13%	90%	5%	5%
apprehensive	8%	17%	75%	5%	45%	50%
concerned	67%	21%	13%	95%	0%	5%
uneasy	0%	29%	71%	5%	20%	75%
upset	0%	17%	83%	0%	0%	100%
worried	79%	17%	4%	70%	20%	10%
excused	21%	33%	46%	25%	25%	50%
legitimate	42%	38%	21%	16%	32%	53%
justified	79%	17%	4%	90%	5%	5%
vindicated	4%	25%	71%	10%	5%	85%
warranted	13%	30%	57%	60%	15%	25%
chances	33%	33%	33%	50%	17%	33%
probability	33%	33%	33%	21%	26%	53%
risk	67%	17%	17%	78%	22%	0%
be compassionate to	5%	9%	86%	5%	11%	84%
be disposed towards	5%	32%	64%	0%	6%	94%
be in favour of	18%	23%	59%	5%	32%	63%
be kind to	0%	19%	81%	5%	11%	84%
support	62%	19%	19%	85%	15%	0%

	Non-native (n=24)			Native (n=20)		
	Very and Completely	Moderately	Not at all and Slightly	Very and Completely	Moderately	Not at all and Slightly
be supportive of	64%	27%	9%	70%	20%	10%
be sympathetic to	32%	23%	45%	45%	25%	30%
sympathize with	45%	27%	27%	30%	55%	15%
abridging	4%	21%	75%	0%	0%	100%
controling	25%	33%	42%	5%	21%	74%
curbing	4%	22%	74%	26%	16%	58%
curtailing	25%	29%	46%	21%	26%	53%
cutting back	55%	14%	32%	28%	28%	44%
limiting	74%	13%	13%	84%	11%	5%
restricting	70%	26%	4%	89%	11%	0%
assembling	5%	45%	50%	11%	16%	74%
collecting	86%	14%	0%	100%	0%	0%
garnering	0%	32%	68%	5%	5%	89%
gathering	68%	32%	0%	79%	11%	11%
pulling together	15%	25%	60%	16%	21%	63%
demanded money for the return of hostages	64%	18%	18%	21%	16%	63%
demanded money for the return of sureties	5%	23%	73%	0%	5%	95%
ransom money for hostages	59%	27%	14%	95%	5%	0%
ransom money for sureties	5%	27%	68%	0%	5%	95%
inclined	64%	27%	9%	42%	21%	37%
prone	68%	32%	0%	53%	16%	32%
advance	5%	18%	77%	0%	6%	94%
boost	23%	36%	41%	6%	6%	89%
encourage	59%	27%	14%	74%	16%	11%
further	9%	32%	59%	0%	17%	83%
promote	68%	23%	9%	79%	11%	11%
afraid	26%	26%	48%	35%	18%	47%
concerned	74%	22%	4%	94%	6%	0%
preoccupied	35%	26%	39%	6%	12%	82%
solicitous	0%	22%	78%	0%	0%	100%
troubled	13%	48%	39%	41%	35%	24%
worried	74%	13%	13%	65%	24%	12%

Note: The shaded cells are the one that are the largest in their rows

Also in the PEW case, there are, of course, differences between native and non-native speakers. First, the least frequent wording, 'menace of terrorism', is viewed more favourably by non-native speakers than native speakers, who prefer the more frequent 'threat of terrorism' (P1). Second, 'worried' was better evaluated by non-native

speakers, while ‘concerned’ was better evaluated by native speakers (P2). Third, ‘justified’, ‘warranted’ and vindicated’ received a better evaluation from native speakers, while ‘legitimate’ received more positive evaluations from non-native speakers (P3). Fourth, non-native speakers liked the original wording, ‘probability’, more than native speakers, who demonstrated more approval for ‘chances’ and ‘risk’ (P4.1). Fifth, native speakers more strongly endorsed both ‘supportive’ wordings and also ‘sympathetic to’, while the wordings ‘sympathise with’ and ‘in favour of’ were better evaluated by non-native speakers (P4.2 and P4.3). This is not surprising, as ‘sympathetic’ is a word that has a different meaning in some other languages (i.e., so called ‘false friend’ as explained in Chapter 2). Sixth, ‘restricting’, ‘limiting’ and ‘curbing’ are liked more by native speakers, while ‘controlling’ and ‘cutting back’ received more endorsement from non-native speakers (P4.4). Seventh, all wordings in the ‘collecting information’ synset were better evaluated by native speakers than non-native speakers – there is no wording that would be better evaluated by a non-native speaker (P4.5). Eighth, non-native speakers better evaluated the wording ‘demanded money’, while ‘ransom money’ received more endorsement by native speakers (P5). Ninth, both ‘inclined’ and ‘prone’ were better evaluated by non-native speakers (P6). Tenth, ‘encourage’ and ‘promote’ were seen as more favourable by native speakers, while the wordings ‘boost’, ‘further’ and ‘advance’ received a better evaluation by non-native speakers (P7). Finally, ‘concerned’, ‘troubled’ and ‘afraid’ were better rated by native speakers, while ‘preoccupied’ and ‘worried’ gained more endorsement from non-native speakers (P8).

4.4 Summary and discussion

Here we summarise the detailed evaluations presented in previous sections of this chapter, focusing on comparisons of the results based on the two approaches. This was also our key research challenge, as we study the extent to which the corpora approach can replace (within this specific context) the expert evaluations.

4.4.1 The summary comparisons for the Wageindicator case

In sum, the expert evaluations and wording frequencies match in five out of the nine cases in the Wageindicator questionnaire. In the remaining four cases, the experts would

choose differently than the frequencies suggested. Table 4.12 summarises wording frequencies (from Table 4.3) and median evaluations.

Table 4.12: Summary comparisons of Wageindicator: corpora approach vs expert evaluations

Original wording	String frequencies (enTenTen)	Expert evaluation – median
W1. Kind of contract	Type (2092) > Kind (954) > Form (872) > …	Type (5) > Kind (4) > …
W2. Wholly owned	Wholly (16383) > All (927) > Completely (287) >…	Completely (4), Entirely (4) > Totally (3), Wholly (3)
W3. Laid down in the contract	Made (48) > Set (40) > Established (38) > …	Set (4) >Established (3) > …
W4. Unforeseen problems	Unexpected (1278) > Unforeseen (1004) > …	Unexpected (5) > Unforeseen (4), Unanticipated (4)
W5.1. Sufficient energy	Enough (9238) > Sufficient (1645) > Adequate (577)	Enough (4), Sufficient (4) < Adequate (3)
W5.2. Sufficient support	Enough (3894) > Adequate (2383) > Sufficient (1336)	
W6. Sufficiently varied	Enough (419) > Sufficiently (55) > Adequately (1)	Sufficiently (4) > Enough (3), Adequate (3)
W7. Good state of repair	Condition (53450) > State of repair (258)	Conditions (4) > State of repair (3)
W8. Staffing levels are sufficient	Adequate (8) > Enough (7) > Sufficient	Enough (4) > Adequate (3), Sufficient (3)

Note: Shaded wordings are those for which wording frequencies and expert evaluations match.

In the Wageindicator evaluation, there were four cases where the expert evaluation and corpora did not match. Further investigation or evaluations with different methods will be needed before making a final decision regarding the most appropriate wording. Of course, the decision can also be made rather arbitrarily. Let us observe these four cases:

- First, although wording frequencies are most favourable for the wording 'wholly owned', most experts considered 'completely owned' as the best wording, among both non-native and native speakers, the latter even more strongly. However, as some experts explained, although it might not be known to every respondent, 'wholly owned' is an established term commonly used in the context of business. Moreover, its frequency is two digits higher than for the other alternatives, which is a strong difference, while the differences in expert evaluations are quite weak (W2).

- Second, although its frequency is not so much higher than for the other alternatives, ‘made in the contract’ is the most frequent wording according to the enTenTen corpus. However, the differences in string frequencies are quite small, so this is a very weak criterion. Moreover, the single frequency for ‘set’ is higher than for ‘made’, and ‘made in the contract’ is the least appropriate according to experts, who would prefer to use ‘set in the contract’ – especially native speakers, while non-native speakers favour ‘made in the contract’ more. In addition, some experts suggested other alternatives that should be further explored; one in particular suggested using ‘set on’ or ‘set by’. (W3).
- Third, although ‘varied enough’ has the highest frequency and is considered completely appropriate by several experts, most of them selected the original wording, ‘sufficiently varied’. However, that string has a much lower frequency. Moreover, some experts suggested alternative wordings that could be considered in further evaluations (W6).
- Fourth, although ‘staffing levels are adequate’ is the best wording according to the corpora, the frequencies are very low and the decision should not be based on them. Moreover, experts put ‘there is enough’ as their first choice. However, with native speakers only, ‘adequate’ would be selected. It appears that both the corpora and experts are quite weak in this case, so it is difficult to make any decision (W8).

In the remaining five cases, corpora frequencies and expert evaluations fully match but the matches vary in strength. For example, according to both methods, the best choice is ‘type of employment’ (W1), but ‘kind’ is not far behind. Similarly, for ‘unexpected problems’, the string frequency (W4) wording is high and has the highest median evaluation, but some other alternatives are also quite high. Next, ‘enough’ has a higher frequency both in the context ‘enough energy’ and ‘enough support’. The experts, however, gave it a similar evaluation to the original wordings, ‘sufficient energy’ and ‘sufficient support’. Moreover, some experts exposed important problems with the wording ‘enough’ in this context. Finally, ‘good conditions’ has both a higher frequency and a better evaluation; however, the latter might change if the expert panel was composed of only native speakers (W7).

4.4.2 The summary comparisons for the PEW case

In sum, with the PEW questionnaire, seven out of the 11 cases matched in both approaches, while for the remaining four, differences occurred. Table 4.13 compares results of the corpora frequencies approach and median expert evaluations.

Table 4.13: Summary comparisons of PEW: corpora approach vs expert evaluations

Original wording	String frequencies (enTenTen)	Expert evaluation – median
P1. Threat of terrorism	Threat (2209) > Menace (126) > Danger (116)	Threat (5) > Danger (3) > …
P2. How worried	Upset (1696) > Concerned (963) > Worried (824) > …	Concerned (4), Worried (4), > Apprehensive (2), Uneasy (2) > …
P3. Ever justified	Justified (227) > Legitimate (33) > Warranted (31) > …	Justified (4) > Excused (3), Legitimate (3), Warranted (3)
P4.1 Chances of attack	Risk (483) > Chances (49) > Probability (28)	Risk (4) > Chances (3), Probability (3)
P4.2 Sympathetic to terrorists	Support (458) >Sympathetic to (18) > Sympathize with (16) > …	Support (4), Be supportive of (4) > Be sympathetic to (3), Sympathize with (3) > …
P4.3 Restricting liberties	Restricting (7) > Curtailing (6) > …	Limiting (4), Restricting (4) > Cutting back (3) > …
P4.4 Collecting information	Gathering (7542) > Collecting (4417) > Assembling (83) > …	Collecting (4), Gathering (4) > Assembling (2), Pulling together (2) > …
P5. Ransom money for hostages	Ransom for hostages (9) > Ransom money for hostages (1) > …	Ransom money (4) > Demanded money (3); Hostages (4) < Sureties (1)
P6. Prone to violence	Prone (452) > Inclined (39)	Prone (3), Inclined (3)
P7. Encourage violence	Promote (1002) > Further (922) > Encourage (517) > …	Encourage (4), Promote (4) > Boost (2), Further (2) > …
P8. Concerned about extremism	Concerned (3) > … (0)	Concerned (4), Worried (4) > Troubled (3) > …

Note: Shaded wordings are those for which wording frequencies and expert evaluations match.

Let us observe in detail the four discrepant cases.

- First, although enTenTen indicated ‘upset’ as the best option, experts preferred the original wording ‘worried’ or ‘concerned’, especially non-native speakers. Moreover, ‘concerned’ and ‘worried’ are not far behind in terms of corpora frequencies (P2).
- Second, even though, formally, ‘restricting liberties’ has a higher corpora frequency than other alternatives, they are actually all very low – in practice,

such a low frequency does not mean much, so this criterion is irrelevant in this case. Based on expert evaluations, the best alternative is either ‘limiting’ or ‘restricting’, but ‘limiting liberties’ is slightly better evaluated by experts, especially native speakers (P4.4).

- Third, ‘gathering information’ is more frequent than ‘collecting information’ according to the text corpora; however, more experts preferred ‘collecting’, in particular native speakers. Considering the median evaluation, both ‘collecting’ and ‘gathering’ have about the same rate, so the difference here is not very serious (P4.5).
- Fourth, ‘promote violence’ is the most frequent but ‘encourage violence’ is the one preferred by experts, especially native speakers. However, the differences in string frequencies are not very high. Moreover, ‘encourage’ and ‘promote’ have about the same median evaluation. So, even in this case, the difference is very small and not serious. (P7).

In any case, further evaluation with different methods would be needed to make a final decision between most of these pairs of words.

In the remaining seven cases, corpora frequencies data and expert evaluations matched but there are differences in strength. One of the strongest is probably ‘threat of terrorism’ (P1), which is much higher and much better evaluated than its alternatives. Similarly, ‘ever justified’ (P3) is also a strong lead, according to both the corpora and expert reviews. Next, ‘risk of attack’ (P4.1) is much more frequent than its alternatives and also better evaluated. Then, ‘support’ (P4.2 and P4.3) also has a much higher frequency than its alternatives but is a less strong lead according to expert reviews, where ‘supportive of’ has the same median value. In contrast, ‘ransom money for hostages’ (P5) is not so strong in terms of corpora frequencies but is a clear favourite of experts. Moreover, ‘prone to violence’ is a strong lead according to corpora frequencies but much less strong according to expert evaluations, where it received the same median evaluation as ‘inclined to violence’ and for both it is quite low. Finally, ‘concerned about extremism’ (P8) is a very weak winner as its frequency is very low, but it is the only one above zero and has the same median evaluation as ‘worried about extremism’.

In addition, for cases P3, P4.1, P.4.2 and P4.3, some experts exposed some shortcomings in the wordings ‘ever justified’, ‘risk of attack’ and ‘support terrorists’. Thus, further evaluations with different methods are also needed for these cases.

4.4.3 Overall summary

We demonstrated that text corpora can be useful in the process of question pretesting, at least when string frequencies are large enough. The latter can be a limitation in some cases (i.e., in around one-quarter of the cases in this study).

The results show that in 12 out of the evaluated 20 cases, we would select the same wording with both approaches. In a further five cases, the differences were actually very small, while in only three cases were more substantive differences found.

The corresponding reasons are rooted in various language specifics, which would require further study and additional procedures. For instance, the alternative wordings might have slightly different meanings, and they were not synonyms with fully equivalent meanings. In part, we examined that in Chapter 5 where the same questions from the PEW study were used in cognitive interviews.

Overall, the results indicate that an approach based on the text corpora – which is semi-automated and inexpensive – can to a considerable extent replace lengthy and resource-demanding expert evaluations. Of course, all of this holds for a specific problem of selection between alternative wordings, while expert evaluations may also have some other benefits.

With respect to the discrepancies, a similar dilemma as in Graesser et al. (2000) and Olson (2010), who compared the QUAID tool with expert evaluations, appears: which approach is closer to the true value (i.e., the right selection) – the approach based on linguistic resources or the expert evaluations? There is no uniform response to this question, although it is generally true that the corpora approach – relying only on wording frequencies – might not include some other aspects of the language. In any case, further comparisons with other methods are needed to determine the answer, which will be done in the final conclusions of the dissertation, after examining the same survey questions with cognitive interviews (Chapter 5) and in a field study (Chapter 6).

We may also add that sometimes there were considerable variations between experts, in particular between experts who are native speakers and those who are non-native speakers. For instance, the word ‘sympathetic’ was better evaluated by native speakers, while non-native speakers would avoid it because of its different connotations in their native languages. As some experts mentioned in their feedback, non-native speakers might not have the right skills to make judgements on terminology issues due to lack of knowledge. However, the development of an English questionnaire by a non-native English speaker is a common situation. Thus, it was important to include them in question pre-testing. The differences between native and non-native speakers are also examined in the cognitive interviews study presented in Chapter 5.

5 Cognitive interviews and text corpora approach

In this chapter, we evaluate a selection of question items and alternative wordings from the PEW questionnaire with cognitive interviews using two techniques: paraphrasing and definitions. The cognitive interviews approach is introduced in Section 2 but here we give more details on the two techniques.

Paraphrasing is a sub-type of verbal probing where the respondent is required to repeat the question with their own words (Lessler and Forsythe 1996; Snijkers 2002; Willis 2005). It allows us to find out if the respondents correctly understand the question, what words are difficult to them, and available suggestions for rewording the question. However, the participants often do not understand their task or find it difficult to think of alternative wordings, which makes this technique not very informative (Snijkers 2002). In particular, it might be problematic to those less educated, with lower cognitive skill and a narrower vocabulary. Thus, instead of literally paraphrasing full questions, Willis (2005) recommends restricting the task to only checking the understanding of specific concepts.

The other cognitive interviewing technique we used required respondents to define a certain concept. According to Mohorko (2015), giving definitions is a technique similar to paraphrasing but is more reliable and gives better results. Although this technique is not very common and is only mentioned in some survey methodology textbooks, as a procedure 'in which respondents provide definitions for key terms in question' (Groves et al. 2009, 264), Mohorko (2015) considers it a form of cognitive interview. However, a possible limitation of this technique is that the participant might only focus on the concept without considering the context of the question. Thus, in cases where there are several potentially problematic concepts, the use of paraphrasing is recommended.

In our study, we used cognitive interviews to compare how respondents understand wording alternatives with the same meaning but which differ in their wording frequencies. Do their definitions and paraphrases reflect our hypothesis that low-frequency wordings are more difficult to comprehend than high-frequency wordings?

In the proceeding sections, we first observe how often respondents who are presented with a low-frequency wording use its high-frequency synonym (or other alternative

wording) when trying to define or paraphrase the term. Second, we also count the number of different definitions and paraphrases given, assuming that there will be a lower variation of definitions and paraphrases for high-frequency wordings as it is clearer to the respondents what they mean. Third, we compare the responses of native and non-native speakers, assuming that the latter will have more difficulties understanding low-frequency wordings. In particular, as discussed in Section 4, certain words such as 'sympathetic' (and other so-called 'false friends') might have different connotations to non-native speakers.

5.1 Methodology

A selection of 13 items from the PEW questionnaire was made, and for each of them two wording alternatives were tested: one with a lower wording frequency and one with a higher wording frequency. Respondents were randomly allocated to either the original version of the PEW questionnaire or the changed version, which replaced the original wording either with a less frequent wording (LF in Table 5.1) (items P0, P1, P2, P3, P6 and P8) or with a more frequent wording (HF in Table 5.1) (items P4.1–P4.5, P5 and P7).

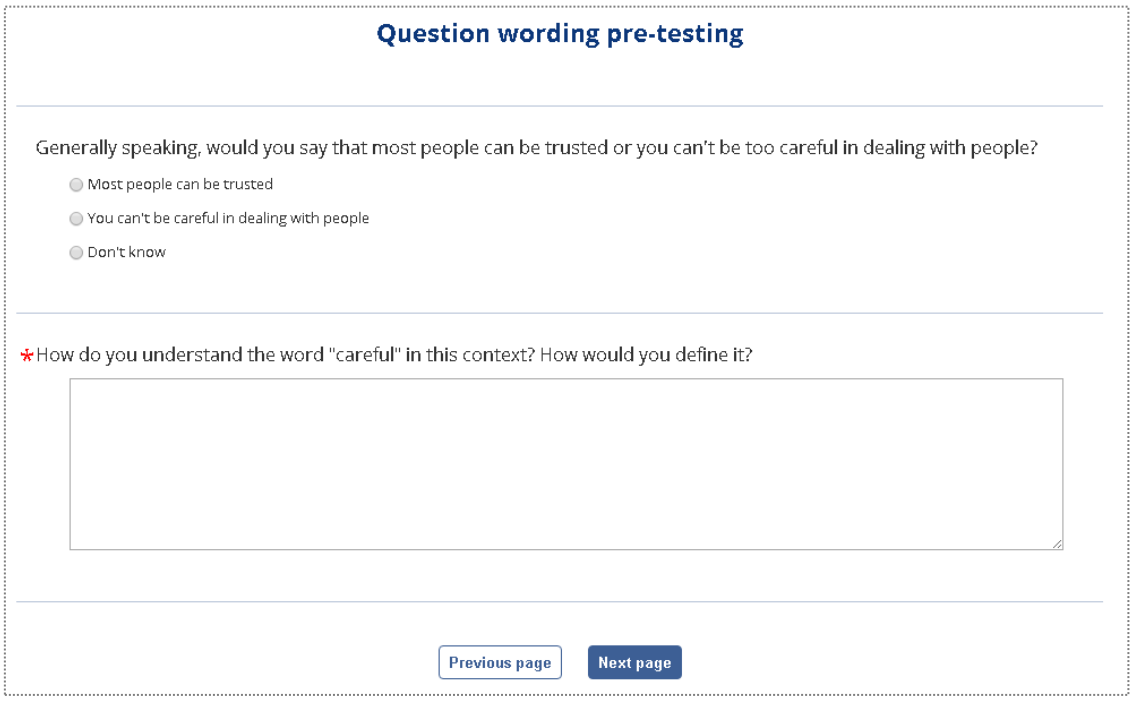
Each question was followed by a probing question that required the participant either to paraphrase the question or to define a specific wording within the question – for each question, we selected the technique we considered most relevant for the task. The definitions technique was used when there was a clear term in the question that might be problematic (items P0, P2, P3 and P8), while paraphrasing was used when there were more than one difficult term in the question or where one problematic term could lead to the wrong understanding of the complete question (items P1, P4.1–P4.5, P5, P6 and P7).

Table 5.1 presents the selected items and shows which wording alternatives and cognitive interviewing techniques were used.

Table 5.1: Cognitive interviewing techniques used for selected items and wording alternatives

Item evaluated (Original version)	Changed version	CI technique
P0. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	cautious (LFW)	define
P1. In general, how well do you think the U. S. government is doing in reducing the threat of terrorism?	menace (LFW)	paraphrase
P2. How worried are you that there will soon be another terrorist attack in the United states?	apprehensive (LFW)	define
P3. Do you think the use of torture against suspected terrorists in order to gain important information can ever be justified?	vindicated (LFW)	define
P4. To what extent do you agree or disagree with the following statements? Please select an answer for each statement.	-	-
P4.1. I often worry about the chances of a nuclear attack by terrorists.	risk (HFW)	paraphrase
P4.2. Freedom of speech should not extend to groups that are sympathetic to terrorists.	support (HFW)	paraphrase
P4.3. The police should be allowed to search the houses of people who might be sympathetic to terrorists without a court order.	support (HFW)	paraphrase
P4.4. The government anti-terrorism policies have gone too far in restricting the average person's civil liberties.	limiting (HFW)	paraphrase
P4.5. I am concerned that the government is collecting too much information about people like me.	gathering (HFW)	paraphrase
P5. As you may know, the United States government has a policy that it NEVER pays ransom money for hostages held by terrorist groups. What is your opinion about this policy?	demanded (HFW)	paraphrase
P6. Which statement comes closer to your own views even if neither is exactly right? Some religions are more prone to violence than others. All religions are about the same when it comes to violence.	inclined (LFW)	paraphrase
P7. Which statement comes closer to your own views even if neither is exactly right? The Islamic religion is more likely to encourage violence among its believers. The Islamic religion does not encourage violence more than other.	promote (HFW)	paraphrase
P8. How concerned, if at all, are you about Islamic extremism around the world these days?	preoccupied (LFW)	define

Figure 5.1: Preview for item P0



When we were interested in the definition, the probe question asked: *'How do you understand the word X in this context? How would you define it?'* For paraphrasing, the wording of the probing question was: *'Could you paraphrase/restate this statement with your own words?'*

Figure 5.1 shows a screenshot of a probing question as an example.

The cognitive interviews were not carried out in person, but an online survey was used to collect responses. Participants were recruited using the crowdsourcing platform Prolific Academic (https://prolific.ac/). Prolific Academic is a platform similar to Amazon Mechanical Turk (https://www.mturk.com), but it is adapted to the needs of academic users. It enables researchers to recruit reliable, on-demand respondents to participant in various studies and experiments. We may add that the researcher (i.e., the user of the Prolific Academic service) has to monetarily compensate participants for

their time with at least 5 GBP per hour (in our case, participants received 1.25 GBP to fill out our questionnaire due to the estimated completion time of 15 minutes). Before approving the award, data quality can be checked. The platform also offers flexible pre-screening where users can choose from a range of demographics to recruit participants.

We used Prolific Academic to run two studies using the same questionnaire. About 60 participants were randomly allocated to one of the two study versions. Only participants who were residents of the United States and were older than 18 years were eligible to participate. The first study was carried out in September 2015, and 60 participants were invited from a pool of 7,756 eligible participants. Of the 60 participants who submitted complete responses, 32 were randomly allocated to the first version and 28 to the second version. However, we subsequently removed one respondent from the first version as his responses were not related to what we asked and not useful for our research.

In February 2016, Prolific Academic introduced new pre-screening filters for participants, including their native language. Thus, we conducted another study but on a slightly different population – all participants who indicated English as their first language were excluded so that it included only those users who listed another language as their first language. Since this significantly reduced the participant pool to only about 300 respondents, we expanded the area of participation to Canada. Thus, in total, there were 471 eligible participants, from which 60 participated in the study. Of the 63 participants who submitted complete responses, 30 were randomly allocated to the first version and 33 to the second version.

In total, 61 evaluated the original version of the questionnaire (31 native and 30 non-native), while 61 evaluated the changed version (28 native and 33 non-native). This design also enabled the comparison of the two samples of respondents: 59 native speakers and 63 non-native speakers.

5.2 Results

We assigned codes to answers in order to summarise the wording used as an alternative to the wording presented in the questionnaire. One answer could be assigned to more than one code. For instance, if somebody listed both ‘cautious’ and ‘skeptical’ as a

definition for ‘careful’, that answer received two codes. Thus, the sum of all responses for a group of respondents is not 100%.

For items that used the definition technique (P0, P2, P3 and P8), the corresponding code was usually very straightforward, and we could just copy what the participant wrote; while for paraphrasing items (P1, P4.1–P4.5, P5, P6 and P7), it was often more difficult to summarise, identify and separate the alternative wording.

In this section, we first summarise the results for the 13 items. We merged answers for both samples (native and non-native speakers), so we are presenting answers for 122 participants, of whom 61 were randomly allocated to the original version and 61 to the changed version, which had six wordings with lower frequencies (LFW) and seven with higher frequencies (HFW). For each case, we also computed the total number of different definitions/paraphrases which might be an indicator of word ambiguity; we also checked if this was reflected in the number of senses of a certain word in WordNet. Besides analysing the most common definitions and paraphrases, we also summarised the differences in responses between native and non-native speakers. Finally, we also provide an overall analysis of the results.

5.2.1 Analysis for individual items

The tables in this section display codes (i.e., coded responses from participants) that appeared at least twice in total. Other answers are merged in the ‘other’ category, which is the sum of all codes that were only mentioned by one participant. In addition, we compute the total number of different responses for a certain item. A high number could indicate that a wording is ambiguous. However, our main focus is on interpreting which answer appears more often in each condition.

In addition, for each item, we also interpret the differences in answers between respondents who responded in the general sample (native and non-native speakers) and those in the non-native-only sample.

5.2.1.1 Defining ‘careful’ and ‘cautious’ (P0)

Table 5.2 presents the results for item P0, where half of the participants were asked to define the word ‘careful’ (HFW – high frequency wording condition) and half the less frequent ‘cautious’ (LFW – low frequency wording condition) in the context of

'Generally speaking, would you say that most people can be trusted or that you can't be too careful/cautious in dealing with people?'

Table 5.2: Frequencies of definitions reported by participants in two wording conditions for item P0 (HWF – high frequency condition; LFW – low frequency condition)

How do you understand the word [careful/cautious] in this context? (P0)	Wording displayed to respondents (merged samples)					
	Careful (HFW) n = 61		Cautious (LFW) n = 61		Total n = 122	
careful	*1	*2%	26	43%	27	22%
cautious	22	36%	0	0%	22	18%
non trusting (easily); hesitate in trusting; examine or get to know or take time before trusting; don't be quick to trust; trust must be earned; not being trusting; distrust	9	15%	5	8%	14	11%
wary	3	5%	6	10%	9	7%
aware	3	5%	5	8%	8	7%
not letting your guard down, keep guard up, guard the fences, guarded, on-guard	4	7%	3	5%	7	6%
safety, safe	3	5%	4	7%	7	6%
weary	3	5%	2	3%	5	4%
avoid taking risk; being risk aware; not taking risks	1	2%	2	3%	3	2%
sceptical/sceptical	2	3%	1	2%	3	2%
avoid danger/getting lied to or being treated badly	1	2%	1	2%	2	2%
defence from the unknown; defensive, less open	1	2%	1	2%	2	2%
not gullible	1	2%	1	2%	2	2%
trustworthy	2	3%	0	0%	2	2%
to watch what you say, being watchful	1	2%	1	2%	2	2%
too trusting	1	2%	1	2%	2	2%
(other – only one mention)	23	38%	13	21%	36	30%
Total number of different definitions	*39*		*27*		*52*	

**There was a respondent assigned to the 'careful' condition that provided the definition 'careful' which is not a valid response in this case.*

In total, there were 52 different definitions given. Within the 'careful' definition, there were 39 different wordings; within the 'cautious' condition, there were 27 different

wordings. A possible explanation is that ‘cautious’ is a less ambiguous wording than ‘careful’. In fact, ‘careful’ has more different senses (5) in WordNet than ‘cautious’ (2).

Of the participants, 36% who were asked to define ‘careful’ defined it as ‘cautious’, and 43% of those asked to define ‘cautious’ defined it as ‘careful’. Apparently, ‘careful’ is a word that more easily comes to mind than ‘cautious’. Other wordings were most frequently related to trusting, especially among those that were in the ‘careful’ condition; while for the ‘cautious’ condition, ‘wary’ and ‘aware’ were two wordings that appeared more often than in the first condition.

In addition, we also checked differences for the two groups of respondents, the general sample, and non-native speakers. The latter used ‘cautious’ (14%) as a definition less often than the general sample (24%). Instead, they used ‘not trusting’ and its variations (17%) more commonly than the general sample (7%).

5.2.1.2 Paraphrasing ‘threat’ and ‘menace’ (P1)

Table 5.3 presents the results for item P1, where participants were asked to paraphrase the question *‘In general, how well do you think the United States government is doing in reducing the threat of terrorism?*’ For half of the sample, the less frequent wording ‘menace’ was used instead of ‘threat’. In their responses, instead of providing any alternatives for the two single wordings, most participants paraphrased the full question, and within it, the string ‘reducing the threat of terrorism’ or ‘reducing the menace of terrorism’. Thus, our codes are based on what alternatives they provided for that string.

Table 5.3: Frequencies of paraphrases reported by participants in two wording conditions for item P1

Could you paraphrase/restate this question with your own words? (P1)	Wording displayed to respondents (merged samples)					
	Threat (HFW) n = 61		Menace (LFW) n = 61		Total n = 122	
reduce/reducing the threat(s)/ terrorism/threat level/terrorist attack/troubles that terrorism is causing/impacts of terrorism/the number of terrorist attacks/amount of terrorist activity	14	23%	14	23%	28	23%
prevent/preventing the threat/(the risk of) terrorism/ (potential) terrorist attacks/terrorism-related events/the	8	13%	6	10%	14	11%

Could you paraphrase/restate this question with your own words? (P1)	Wording displayed to respondents (merged samples)					
	Threat (HFW) n = 61		Menace (LFW) n = 61		Total n = 122	
harm/the spread of terrorism						
deal/dealing with terrorism/ the menace/threat of terrorism	4	7%	4	7%	8	7%
combat terrorism; combating terrorism/terroristic threats	3	5%	4	7%	7	6%
war on terror/terrorism	3	5%	3	5%	6	5%
counter terrorism/countering terrorist threats/counterterrorism efforts/counter-terrorism measures	0	0%	4	7%	4	3%
stop/stopping terrorism/the danger of terrorism/terrorist attacks	0	0%	3	5%	3	2%
protects people/protecting (its) people	1	2%	2	3%	3	2%
decreasing terrorism/terrorist attacks	2	3%	0	0%	2	2%
fight against terrorism	1	2%	1	2%	2	2%
lessen the risk/likelihood of terrorism	1	2%	1	2%	2	2%
lowering the threat	1	2%	1	2%	2	2%
handling terrorist threats/threat of terrorism	0	0%	2	3%	2	2%
policies to stop terrorism/policy regarding terrorism	0	0%	2	3%	2	2%
NA	1	2%	1	2%	2	23%
(other – only one mention)	21	34%	16	26%	37	30%
Total number of different paraphrases	*31*		*29*		*51*	

In total, participants provided 51 different responses and the number is about 30 for both conditions, even though ‘threat’ has more senses in WordNet (4) than ‘menace’ (2).

The most often type of paraphrasing in both conditions was keeping the wording ‘reducing’ (or simplifying it to ‘reduce’) and changing ‘threat’ or ‘menace’ to various expressions, some including the word ‘threat’ (i.e., ‘threat level’) but most without it and instead focusing only on ‘terrorism’ (i.e., ‘terrorist attack’, ‘troubles that terrorism is causing’, ‘impacts of terrorism’, ‘the number of terrorist attacks’, ‘amount of terrorism’). Similarly, there were some paraphrases that replaced ‘reduce’ with ‘prevent/preventing’, ‘deal/dealing’, etc. In most cases, there were no differences (or only minor ones) between the two conditions, with the exception of ‘counter-terrorism’

(and its variations) and 'stop/stopping terrorism' (and its variations), which only appeared in the 'menace' condition.

In addition, we also checked the differences between the two samples. On the one hand, 'preventing' was mostly used by the general sample (17% vs. 7% among native speakers) and 'decreasing', 'lowering' and 'handling' were used exclusively by the general sample. On the other hand, the wording 'dealing' was more common among non-native speakers (10% vs. 3% in the general sample) and there were some wordings exclusively used by the non-native sample, i.e., 'counter-terrorism' (and its variations), 'protects/protecting people' and 'policies to stop terrorism/regarding terrorism'.

5.2.1.3 Defining 'worried' and 'apprehensive' (P2)

Table 5.4 presents the results for item P2, where half of the participants were asked to define the word 'worried' and half the less frequent 'apprehensive' in the context of '*How worried are you that there will soon be another terrorist attack in the United States?*'

Table 5.4: Frequencies of definitions reported by participants in two wording conditions for item P2

How do you understand the word [worried/apprehensive] in this context? (P2)	Wording displayed to respondents (merged samples)					
	Worried (HFW) n = 61		Apprehensive (LFW) n = 61		Total n = 122	
fearful; fearfulness; (constant) fear	7	11%	18	30%	25	20%
(highly) concerned; (feeling) concern	19	31%	5	8%	24	20%
anxious; feeling anxiety; overrun by anxious thoughts	10	16%	12	20%	22	18%
worried; sense of worry	3	5%	13	21%	16	13%
scared	4	7%	7	11%	11	9%
thinking about it/another attack is likely to happen; thinking something will happen/it is likely to happen	8	13%	1	2%	9	7%
afraid	2	3%	4	7%	6	5%
expectant; expecting; expect the unexpected	4	7%	2	3%	6	5%
anticipating (something); anticipate (the probability)	3	5%	2	3%	5	4%

How do you understand the word [worried/apprehensive] in this context? (P2)	Wording displayed to respondents (merged samples)					
	Worried (HFW) n = 61		Apprehensive (LFW) n = 61		Total n = 122	
nervous	2	3%	2	3%	4	3%
believing there is a possibility; believing there will be	2	3%	1	2%	3	2%
dreading; sense of dread	1	2%	2	3%	3	2%
preoccupation/preoccupied	2	3%	1	2%	3	2%
waiting an attack; expecting an attack; chances of another attack	0	0%	3	5%	3	2%
wary	0	0%	3	5%	3	2%
bothered	1	2%	1	2%	2	2%
likelihood, how likely	1	2%	1	2%	2	2%
mental energy spent/preparing mentally	1	2%	1	2%	2	2%
on edge	0	0%	2	3%	2	2%
NA	0	0%	1	2%	1	20%
(other – only one mention)	14	23%	6	10%	20	16%
Total number of different definitions	*24*		*19*		*33*	

In total, responses were listed in 33 different codes. Those in the 'worried' condition gave slightly more different answers (24) than those in the 'apprehensive' condition (19), which might reflect that the first is more ambiguous and the second has a narrower meaning. However, this is not reflected in WordNet as 'apprehensive' has actually one more sense (3) than 'worried' (2).

Only one participant who had to define 'worried' used 'apprehensive' in the definition, while 21% of those with the task of defining 'apprehensive' used 'worried' in the definition. However, for them, the most popular choice was 'fearful' and its variations, which were selected by 30% of participants in the 'apprehensive' condition compared to only 11% of those assigned to the 'worried' condition. It was followed by 'anxious' and its variations (20%), which was quite a popular answer in the other 'worried' condition as well (16%). However, for those in this condition, the most frequent response was 'concerned' and its variations, which were used by 31% of participants compared to only 8% for the other condition. Another popular choice was also 'thinking' (13% compared to only 2% in the other condition).

In addition, we checked differences between the two samples. Those in the non-native sample used the word 'fear' (29%) more often than those in the general sample (14%). On the other hand, those in the general sample used 'concerned' (27%) and 'anxious' (22%) more often than non-native speakers (14% for 'concerned' and 15% for 'anxious'). Moreover, several words were used exclusively by the general group, i.e., 'nervous', 'dreading', 'bothered' and 'on edge'.

5.2.1.4 Defining 'justified' and 'vindicated' (P3)

Table 5.5 presents the results for item P3, where half of the participants were asked to define the word 'justified' and half the less frequent 'vindicated' in the context of '*Do you think the use of torture against suspected terrorists in order to gain important information can ever be justified?*'

Table 5.5: Frequencies of definitions reported by participants in two wording conditions for item P3

How do you understand the word [justified/vindicated] in this context? (P3)	Wording displayed to respondents (merged samples)					
	Justified (HFW) n = 61		Vindicated (LFW) n = 61		Total n = 122	
(being) justified; justifying; justify; justifiable (when there is no doubt); justification	2	3%	35	57%	37	30%
accepted, (morally) acceptable (in circumstances)	13	21%	0	0%	13	11%
moral/morally proper/morally okay/morally right/doing morally	4	7%	2	3%	6	5%
clear; clear/clearing someone blame/ (of) suspicions	2	3%	3	5%	5	4%
excusable; excuse; excused (because of circumstances)	2	3%	3	5%	5	4%
okay; ok (for a greater reason)	3	5%	2	3%	5	4%
worth; worth it/the (ethical) cost	3	5%	2	3%	5	4%
legitimate; legitimized	3	5%	1	2%	4	3%
necessary to use/ for greater good; (understood to be) necessary	3	5%	1	2%	4	3%
condoned	0	0%	3	5%	3	2%
having a (good) reason	2	3%	1	2%	3	2%
allowable; allowed	1	2%	1	2%	2	2%
do whatever without morality factor;	1	2%	1	2%	2	2%

How do you understand the word [justified/vindicated] in this context? (P3)	Wording displayed to respondents (merged samples)					
	Justified (HFW) n = 61		Vindicated (LFW) n = 61		Total n = 122	
moral issues are not a factor						
ethical, doing ethically	2	3%	0	0%	2	2%
just	0	0%	2	3%	2	2%
proven/prove (to be) right	0	0%	2	3%	2	2%
reasoned; reasonable; having a good reason	2	3%	0	0%	2	2%
Other	22	34%	17	26%	39	32%
Total number of different definitions	*37*		*32*		*57*	

In total, 57 different codes were given, with a few more in the 'justified' condition (37) than in the 'vindicated' condition, where most of the participants agreed on what the definition is. In WordNet, both have only one sense, though.

More than half (57%) of the participants who were assigned to the 'vindicated' condition defined it using the word 'justified' and its variations. In contrast, none of those assigned to the 'justified' condition used the word 'vindicated' to define it. Instead, the most popular definitions were 'accepted' and its variations (21%) and 'moral' and its variations, which were much less popular among those in the other condition.

In addition, we compared the general and non-native sample. There were a few answers that were exclusive for one or the other, specifically 'condoned', 'allowable/allowed', 'without morality factor' and 'reasoned/reasonable' only appeared in the 'justified' condition. On the other hand, 'clearing blame/suspicion' and 'ethical/ethically' only appeared in the 'vindicated' condition.

5.2.1.5 Paraphrasing 'chances' and 'risk' (P4.1)

Table 5.6 presents the results for item P4.1, where participants were asked to paraphrase the statement '*I often worry about the risk of a nuclear attack by terrorists.*' For half of the sample, the more frequent wording, 'risk', was used instead of 'chances'. In their responses, participants did not provide almost any alternatives for the two single wordings, but they were paraphrasing the full question, and within it, the string 'chances

of a nuclear attack' or 'risk of a nuclear attack'. Thus, our codes are based on what alternatives they provided for that string.

Table 5.6: Frequencies of paraphrases reported by participants in two wording conditions for item P4.1

Could you paraphrase/restate the statement *'I often worry about the [chances/risk] of a nuclear attack by terrorists'* with your own words? (P4.1)	**Wording displayed to respondents (merged samples)**					
	Chances (LFW) n = 61		**Risk (HFW) n = 61**		**Total n = 122**	
worry about (that)	3	5%	8	13%	11	9%
possibility	3	5%	5	8%	8	7%
risk	0	0%	7	11%	7	6%
real threat/serious threat/threat/threatens my daily life	3	5%	3	5%	6	5%
there could be/there will be	1	2%	5	8%	6	5%
(how) likely	5	8%	2	3%	5	6%
fear/fear frequently/fear that will soon launch/fear the day/frequently fear	0	0%	3	5%	3	2%
highly concerned/concerned/concern/concerns/feel concern	1	2%	2	3%	3	2%
may find some way/may occur/may use	3	5%	0	0%	3	2%
might use/might be/might get	1	2%	2	3%	3	2%
often worried/often worry about	3	5%	0	0%	3	2%
afraid that will be affected/afraid/afraid of getting	2	3%	0	0%	2	2%
apprehensive/regularly apprehensive	1	2%	1	2%	2	2%
chance/chances	2	3%	0	0%	2	2%
frequently	1	2%	1	2%	2	2%
immanent	0	0%	2	3%	2	2%
likelihood	2	3%	0	0%	2	2%
often on my mind/often think about	1	2%	1	2%	2	2%
scared	1	2%	1	2%	2	2%
something I worry about/sometimes worry	0	0%	2	3%	2	2%
always think that are capable/believer that are capable	0	0%	2	3%	2	2%
(other – only one mention)	19	31%	10	16%	29	24%
Total number of different paraphrases	*35*		*26*		*51*	

In total, 51 different paraphrases were given. It is interesting that none of the respondents in the 'chances' condition used 'risk' and vice versa; those in the 'risk' condition did not resort to the wording 'chances' in their paraphrases. Instead, 'worry' is the most popular choice in this condition, while among those in the first it is 'likely' (8%). However, even that one is not a clear winner as the answers are quite varied. This can also be observed from the number of different paraphrases, which is higher in the 'chances' condition (35) than in the 'risk' condition. Apparently, 'chances' is a more ambiguous wording, although the number of senses in WordNet (5) is only one unit higher than for 'risk' (4).

In addition, we compared those in the general sample to those in the non-native sample. There are few differences in their responses, but there are some items that appear exclusively in only one of the two groups. For instance, 'afraid', 'apprehensive', 'chances' and 'frequently' were only given by those in the general sample, while 'imminent', 'likelihood' and 'often on mind' were specific for the non-native sample.

5.2.1.6 Paraphrasing 'sympathetic to' and 'support' (P4.2 and P4.3)

Table 5.7 presents the results for item P4.2, where participants were asked to paraphrase the statement '*Freedom of speech should not extend to groups that are sympathetic to terrorists*'. For half of the sample, the more frequent wording 'support' was used instead of 'sympathetic to'. In their responses, most of the participants did not provide alternatives for the two single wordings, but they were paraphrasing the full question, and within it, the string 'groups that are sympathetic to terrorists' or 'groups that support terrorists'. Thus, our codes are based on what alternatives they provided for those two strings.

Table 5.7: Frequencies of paraphrases reported by participants in two wording conditions for item P4.2

Could you paraphrase the statement *'Freedom of speech should not extend to groups that [are sympathetic to/support] terrorists'* with your own words? (P4.2)	**Wording displayed to respondents (merged samples)**					
	Are sympathetic to (LFW) n = 61		**Support (HFW) n = 61**		**Total n = 122**	
(terrorist) supporter/support/supported/ supporting; supporters (of terrorists); groups that (publicly) support terrorists; those supporting terrorist ideals; people	15	25%	32	52%	47	39%

Could you paraphrase the statement *'Freedom of speech should not extend to groups that [are sympathetic to/support] terrorists'* with your own words? (P4.2)	**Wording displayed to respondents (merged samples)**					
	Are sympathetic to (LFW) n = 61		**Support (HFW) n = 61**		**Total n = 122**	
who support terrorism/ terrorists						
(terrorist) sympathizer(s); sympathizer (of terror);groups that sympathise/are sympathetic/express sympathy to (terrorists); terrorist-sympathetic; views that are sympathetic to (terrorists)	19	31%	1	2%	20	16%
terrorists	1	2%	5	8%	6	5%
(who) side with terrorists; groups on the side of terrorists	5	8%	0	0%	5	4%
agree(s) with terrorists	4	7%	0	0%	4	3%
pro terrorism groups/pro-terrorist	1	2%	1	2%	2	2%
(other – only one mention)	17	28%	12	20%	29	24%
Total number of different paraphrases	*23*		*16*		*35*	

In total, 35 different paraphrases were listed by participants: 23 by those who were randomly assigned to the 'sympathetic' condition, and 16 by those in the 'support' condition, which appeared to be less ambiguous than the first wording. However, it has a higher number of senses in WordNet (11) than 'sympathetic' (6).

Half of the participants in the 'support' condition used the same wording and its variations (i.e., supporters, supporting, etc.) in the provided paraphrase. Similarly, almost one-third of those in the 'sympathetic to' condition used the same wording they were presented with and its variations. On the other hand, a quarter of those in this condition used 'support' in the paraphrase, while 'sympathetic' was used only by one participant in the 'support' condition. Moreover, those in the 'support' condition more often labelled the groups as not only supporters but actual 'terrorists', while the paraphrases 'side with terrorists' and 'agree with terrorists' were listed exclusively by those in the 'sympathetic to' condition.

In addition, we compared the general and non-native speakers sample. The only two notable differences are that the 'pro-terrorism group' wording was only used by those in the general sample, while referring to them simply as terrorists was (with one exception) specific to the non-native sample.

Table 5.8 presents the results for item P4.3, where participants were asked to paraphrase the statement '*The police should be allowed to search the houses of people who might be sympathetic to terrorists without a court order*'. Same as for item P4.2, half of the sample was presented with the more frequent wording 'support' instead of 'sympathetic to'. In their responses, participants did not provide almost any alternatives for the two single wordings, but they were paraphrasing the full question, and within it, the string 'groups that are sympathetic to terrorists' or 'groups that support terrorists'. Thus, our codes are based on what alternatives they provided for that string.

Table 5.8: Frequencies of paraphrases reported by participants in two wording conditions for item P4.3

Could you paraphrase the statement '*The police should be allowed to search the houses of people who might [be sympathetic to/support] terrorists without a court order*' with your own words? (P4.3)	**Wording displayed to respondents (merged samples)**					
	Sympathetic to (LFW) n = 61		**Support (HFW) n = 61**		**Total n = 122**	
(alleged) (terrorist) supporter(s); support/supporting terrorists; those who (might) support terrorist	8	13%	23	38%	31	25%
sympathizers; sympathetic; sympathise; sympathize	18	30%	1	2%	19	16%
suspected helping/ of supporting/ terrorist supporters/terrorist affiliate/ they may be hiding something; suspicious people	4	7%	15	25%	19	16%
suspected terrorist(s); suspects/ suspicion of terrorism; possible suspect of terrorism	4	7%	2	3%	6	5%
(potential) terrorists	0	0%	6	10%	6	5%
aiding terrorist/aid/those that aid	2	3%	1	2%	3	2%
agree (with)	3	5%	0	0%	3	2%
NA	2	3%	9	15%	11	9%
(other – only one mention)	22	36%	3	5%	25	20%
Total number of different paraphrases	*29*		*10*		*33*	

In total, 33 differently coded paraphrases were given again, most of them by those in the 'sympathetic to' condition (29), which appears to be a more ambiguous wording than 'support' (10), even if the latter has actually more different senses in WordNet. In the

latter condition, 38% provided a paraphrase using the wording 'support' and its variations, while only 13% in the first condition used it. Similarly, in the first condition, 30% of participants used the root 'sympathise' and its variations, but in the second condition, only one participant used it. Moreover, those in the 'support' condition were more likely to use the wording 'suspected' and its variations and to also consider the supporters simply as 'terrorists'. The latter did not appear among those in the 'sympathetic to' condition.

In addition, we compared the general sample and non-native speakers' sample. The first sample was more likely to use either 'support' or 'sympathetic' in their paraphrase, while the non-native sample provided more variation in their answers. A paraphrase that was almost exclusive to them was referring to those groups simply as 'terrorists'.

5.2.1.7 Paraphrasing 'restricting' and 'limiting' (P4.4)

Table 5.9 presents the results for item P4.4, where participants were asked to paraphrase the question '*The government's anti-terrorism policies have gone too far in restricting the average person's civil liberties*'. For half of the sample, the more frequent wording 'limiting' was used instead of 'restricting'. In their responses, participants did not provide almost any alternatives for the two single wordings, but they were paraphrasing the full question, and within it, the string 'restricting (…) civil liberties' or 'limiting (…) civil liberties'. Thus, our codes are based on what alternatives they provided for that string.

Table 5.9: Frequencies of paraphrases reported by participants in two wording conditions for item P4.4

Could you paraphrase the statement *'The government anti-terrorism policies have gone too far in [restricting/limiting] the average person's civil liberties'* with your own words? (P4.4)	**Wording displayed to respondents (merged samples)**					
	Restricting (LFW) n = 61		**Limiting (HFW) n = 61**		**Total n = 122**	
restrict; restricting; restricted; restriction	20	33%	2	3%	22	18%
limit; limited; limiting	1	2%	19	31%	20	16%
infringe; infringed; infringing	1	2%	6	10%	7	6%
affected; affecting; affect average citizen's life	1	2%	3	5%	4	3%

Could you paraphrase the statement *'The government anti-terrorism policies have gone too far in [restricting/limiting] the average person's civil liberties'* with your own words? (P4.4)	**Wording displayed to respondents (merged samples)**					
	Restricting (LFW) n = 61		**Limiting (HFW) n = 61**		**Total n = 122**	
(have) gone (too far); gotten out of hand	0	0%	4	7%	4	3%
invade (and restrain); invaded; invasion of privacy	3	5%	0	0%	3	2%
impact; impacted over	2	3%	0	0%	2	2%
overextended its reach; over-extended their authority	2	3%	0	0%	2	2%
overstepped its boundaries/bounds	0	0%	2	3%	2	2%
restrain; restraints	1	2%	1	2%	2	2%
rights taken away	1	2%	1	2%	2	2%
sacrificed	1	2%	1	2%	2	2%
take too much freedom and privacy; taken away freedoms	2	3%	0	0%	2	2%
NA	1	2%	4	7%	5	4%
(other – only one mention)	23	38%	18	30%	41	34%
Total number of different paraphrases	*35*		*28*		*55*	

In total, 55 paraphrases were given, a few more for the less frequent 'restricting' condition (35) compared to the more frequent 'limiting' condition (28). Correspondingly, 'restricting' has one more sense (4) in WordNet than 'limiting' (3).

As for some of the previous cases, one-third of those in the 'restricting' condition used the word 'restrict' and its variations, and almost one-third of those in the 'limiting' condition also used 'limit' and its variations. Only one in the first condition and two in the second condition used the wording from the opposite condition. Moreover, there were also some other differences in responses between the two conditions. Those assigned to 'limiting' were more likely to use 'infringe', 'affect' and 'gone too far', while those responding to the 'restricting' version were particular in their use of the words 'impact' and 'overextended'.

In addition, we compared the two samples and found that there are some paraphrases that were used exclusively by only one of the two groups. For instance, the wordings 'overstepped its boundaries', 'rights taken away' and 'take away freedoms' were only

used by those in the general sample, while non-native speakers were particular in providing the wordings 'impact' and 'sacrificed'.

5.2.1.8 Paraphrasing 'collecting' and 'gathering' (P4.5)

Table 5.10 presents the results for item P4.5, where participants were asked to paraphrase the question '*I am concerned that the government is collecting too much information about people like me*'. For half of the sample, the more frequent wording 'collecting' was used instead of 'gathering'. The codes are focused on what alternatives participants provided, either for the two verbs or for the strings 'collecting information' and 'gathering information'.

Table 5.10: Frequencies of paraphrases reported by participants in two wording conditions for item P4.5

Could you paraphrase/restate the statement *'I am concerned that the government is [collecting/gathering] too much information about people like me'* with your own words? (P4.4)	**Wording displayed to respondents (merged samples)**					
	Collecting (LFW) n = 61		**Gathering (HFW) n = 61**		**Total n = 122**	
collect; (being) collected; collecting; collective; (covert) collection	32	52%	14	23%	46	38%
gather(s); gathered; gathering (too much) information	6	10%	19	31%	25	20%
spies; spying; snooping	6	10%	5	8%	11	9%
monitored; monitoring	2	3%	5	8%	7	6%
surveil; (mass) surveillance	0	0%	5	8%	5	4%
obtaining/obtained too much	0	0%	3	5%	3	2%
getting/having too much information	2	3%	0	0%	2	2%
has access	0	0%	2	3%	2	2%
having information (about me)	1	2%	1	2%	2	2%
having right to information	1	2%	1	2%	2	2%
infringing	1	2%	1	2%	2	2%
interfering with/invading privacy	1	2%	1	2%	2	2%
knows too much	1	2%	1	2%	2	2%
storing	2	3%	0	0%	2	2%
watching	1	2%	1	2%	2	2%
NA	1	2%	1	2%	2	2%
(other – only one mention)	8	13%	9	15%	17	14%

Could you paraphrase/restate the statement *'I am concerned that the government is [collecting/gathering] too much information about people like me'* with your own words? (P4.4)	**Wording displayed to respondents (merged samples)**		
	Collecting (LFW) n = 61	**Gathering (HFW) n = 61**	**Total n = 122**
Total number of different paraphrases	*21*	*25*	*33*

In total, paraphrases were coded into 33 codes, and about the same number were given by respondents in both conditions. Based on the results for other cases, we would have actually expected more variation for 'gathering' which has a higher number of senses in WordNet (9) than 'collecting' (5).

More than half of those in the 'collecting' condition used 'collect' and its variations in their paraphrases and almost one-third of those in the 'gathering' condition used 'gather' and its variations. On the other hand, only 23% of the latter condition used 'collect', and only 10% of those in the first condition used 'gather' in their paraphrases. There are also other differences between the two conditions: only those in the first condition used the wordings 'getting/having too much information' and 'storing', while the wordings 'surveillance' and 'having access to' were only used by those in the second condition.

In addition, we compared the two samples according to language. Several terms were used only in paraphrases of non-native speakers, i.e., 'surveillance', 'obtaining too much information', 'having access to' and 'storing'. On the other hand, the paraphrase 'having the right to information' was only used by two respondents in the general sample.

5.2.1.9 Paraphrasing 'ransom money' and 'demanded money' (P5)

Table 5.11 presents the results for item P5, where participants were asked to paraphrase the question '*As you may know, the United States government has a policy that it NEVER pays ransom money for hostages held by terrorist groups. What is your opinion about this policy?*' For half of the sample, the more frequent wording 'demanded money' was used instead of 'ransom money'. Our codes are based on what alternatives they provided for these two strings.

Table 5.11: Frequencies of paraphrases reported by participants in two wording conditions for item P5

Could you paraphrase/restate this question with your own words? (P5)	Wording displayed to respondents (merged samples)					
	Ransom money (LFW) n = 61		Demanded money (HFW) n = 61		Total n = 122	
ransom/ransoms (money)	36	59%	21	34%	57	47%
negotiate/negotiates/negotiating/negotiation	10	16%	16	26%	26	21%
demand/demanded/demanding money/ money they demanded/demanded (hostage) money; (not) give in to the/their demands	2	3%	15	25%	17	14%
give/giving (them) money (in exchange)	4	7%	0	0%	4	3%
money	2	3%	3	5%	5	4%
pay terrorists	1	2%	1	2%	2	2%
no wording	1	2%	1	2%	2	2%
(other – only one mention)	5	8%	6	10%	11	9%
Total number of different paraphrases	*22*		*24*		*37*	

In total, paraphrases were coded into 37 codes, and about the same number were given by respondents in both conditions, although 'demanded' has more senses in WordNet (5) than 'ransom' (1).

Almost 60% of those in the 'ransom' condition used 'ransom' in their paraphrases, while 'demand' and its variations were used only by 25% of those in the 'demand' condition. On the other hand, 34% of those in this condition selected 'ransom', while 'demand' was only selected by 3% of the respondents in the 'ransom condition'. Moreover, those in the 'demanded' condition more often (26%) used the word 'negotiate' and its variations than those in the 'ransom' condition. Another difference between the two conditions is that the paraphrase 'giving them money' was only used by those randomly assigned to the 'ransom' condition.

In addition, we compared the results for the general and non-native samples. The latter group less often used the wording 'ransom' than the general sample, and there were some paraphrases specific for them, i.e., 'pay terrorists' (without using 'ransom', 'demanded', or something else).

5.2.1.10 Paraphrasing 'prone to' and 'inclined to' (P6)

Table 5.12 presents the results for item P6, where participants were asked to paraphrase the statement '*Some religions are more prone to violence than others*'. For half of the sample, the less frequent wording 'inclined to' was used instead of 'prone to'. When coding answers, we focused on what alternatives respondents provided for the words 'prone' and 'inclined'.

Table 5.12: Frequencies of paraphrases reported by participants in two wording conditions for item P4.6

Could you paraphrase/restate the statement 'Some religions are more [prone/inclined] to violence than others' with your own words? (P6)	Wording displayed to respondents (merged samples)					
	Prone to (HFW) n = 61		Inclined to (LFW) n = 61		Total n = 122	
more violent	6	10%	4	7%	10	8%
more prone/prone	1	2%	8	13%	9	7%
more likely	2	3%	6	10%	8	7%
inclined	0	0%	6	10%	6	5%
encouraged; encourage (violence more)	5	8%	0	0%	5	4%
lead to more	1	2%	3	5%	4	3%
promoting violence; promote	2	3%	2	3%	4	3%
allow (to be)	2	3%	1	2%	3	2%
justify	1	2%	2	3%	3	2%
are more violent; be more violent; being closer to violence	2	3%	0	0%	2	2%
greater propensity towards	0	0%	2	3%	2	2%
incite aggression/more	1	2%	1	2%	2	2%
more eager	1	2%	1	2%	2	2%
more susceptible	2	3%	0	0%	2	2%
more well-known when talking about	0	0%	2	3%	2	2%
(other – only one mention)	28	46%	19	31%	47	39%
Total number of different paraphrases	*40*		*31*		*62*	

There were 62 different paraphrases given by respondents, a few more for the 'prone' condition (40) than for the 'inclined' condition (31), indicating that the latter might be

clearer and less ambiguous. However, this is not reflected in the number of senses in WordNet, which is one unite higher for 'inclined' (3) than for 'prone' (2).

13% of those in the 'inclined' condition used 'prone' and 10% used 'inclined' in their paraphrases, while only one respondent in the 'prone' condition used 'prone' and none used 'inclined'. Instead, they used the wordings 'more violent' and 'encouraged' in their answers.

In addition, we compared responses between the general sample and non-native speakers and found some interesting differences. 'Inclined', '(lead to) more' and 'justify' were used exclusively by the latter, while only the first group used the words 'great propensity towards', 'more eager' and 'more well-known'.

5.2.1.11 Paraphrasing 'encourage' and 'promote' (P7)

Table 5.13 presents the results for item P7, where participants were asked to paraphrase the statement '*The Islamic religion is more likely to encourage violence among its believers*'. For half of the sample, the more frequent wording 'promote' was used instead of 'encourage'. When coding answers, we focused on what alternatives respondents provided for these two words.

Table 5.13: Frequencies of paraphrases reported by participants in two wording conditions for item P7

Could you paraphrase the statement 'The Islamic religion is more likely to [encourage/promote] violence among its believers' with your own words? (P7)	Wording displayed to respondents (merged samples)					
	Encourage (LFW) n = 61		Promote (HFW) n = 61		Total n = 122	
more likely to be	3	5%	9	15%	12	10%
promoted/promotes/promote/greater promotion/tendency to promote	3	5%	9	15%	12	10%
more prone/prone	8	13%	2	3%	10	8%
encourage/encouraged/encouraged to commit/encourages	7	11%	1	2%	8	7%
more likely to promote	0	0%	8	13%	8	7%
more	1	2%	3	5%	4	3%
more common	3	5%	0	0%	3	2%
support/supports	1	2%	2	3%	3	2%

Could you paraphrase the statement 'The Islamic religion is more likely to [encourage/promote] violence among its believers' with your own words? (P7)	Wording displayed to respondents (merged samples)					
	Encourage (LFW) n = 61		Promote (HFW) n = 61		Total n = 122	
advocate/advocates	0	0%	2	3%	2	2%
more likely to incite	1	2%	1	2%	2	2%
more likely to lead	1	2%	1	2%	2	2%
preaches violence/preach violence	2	3%	0	0%	2	2%
spread/spreads	1	2%	1	2%	2	2%
tend to be more/tends to be	2	3%	0	0%	2	2%
NA	2	3%	1	2%	3	2%
(other – only one mention)	24	39%	23	38%	47	39%
Total number of different paraphrases	*37*		*35*		*62*	

Respondents provided 62 different paraphrases, about the same number as those in the 'encourage' and 'promote' condition. In WordNet, however, 'encourage' has les WordNet senses (3) than 'promote' (5).

In the latter condition, 15% used 'promote', while 11% from the former condition used 'encourage'. On the other hand, only 5% of those in the 'encourage' condition used the word 'promote', and only one participant in the 'promote' condition used the word 'encourage'. In the 'encourage' condition, the most popular paraphrase was 'prone', while in the 'promote' condition it was, apart from 'promote', also 'more likely to be' and 'more likely to promote', which were exclusive to this condition. Another paraphrase exclusive to those assigned to the 'promote' condition is 'advocate'. On the other hand, the paraphrases 'more common' and 'preach violence' are exclusive to the 'encourage' condition.

In addition, we compared the general and non-native samples. The first were more likely to use the wordings 'more likely', 'encourage' and 'more likely to promote' in their paraphrases, while non-native speakers were more likely to use 'promote' and its variations. Moreover, the paraphrases 'more common' and 'advocate' were only used by those in the 'encourage' condition, while 'more likely to lead' and 'preach violence' were exclusive to the 'promote' condition.

5.2.1.12 Defining 'concerned' and 'preoccupied' (P8)

Table 5.14 presents the results for item P8, where half of the participants were asked to define the word 'concerned' and half the less frequent 'preoccupied' in the context of *'How concerned, if at all, are you about Islamic extremism around the world these days?'*

Table 5.14: Frequencies of definitions reported by participants in two wording conditions for item P8

How do you understand the word [concerned/preoccupied] in this context? (P8)	Wording displayed to respondents (merged samples)					
	Concerned (HFW) n = 61		Preoccupied (LFW) n = 61		Total n = 122	
worrying about; (constant) worry; worried (about); pre-worried	44	72%	7	11%	51	42%
(often) think: thinking; (often) thinking about; thinking about constantly; thinking about (something) very often; thinking much on something; thought about; thoughtful of it; the extent of thinking about; spend thinking about it	4	7%	18	30%	22	18%
concern; concerned	0	0%	10	16%	10	8%
preoccupied	1	2%	8	13%	9	7%
(be) busy	0	0%	6	10%	6	5%
always on your mind; being unable to get it off your mind; mind busy on; frequently on your mind	0	0%	4	7%	4	3%
anxiety; anxious	4	7%	0	0%	4	3%
affected	0	0%	2	3%	2	2%
afraid	2	3%	0	0%	2	2%
focused heavily; focused on more	0	0%	2	3%	2	2%
informed	0	0%	2	3%	2	2%
involved	0	0%	2	3%	2	2%
scared; scares	2	3%	0	0%	2	2%
NA	7	11%	2	3%	9	7%
(other – only one mention)	9	15%	19	31%	28	23%
Total number of different definitions	*12*		*23*		*34*	

Thirty-four different codes were used to categorise the definitions provided by participants, and substantially more were given by those in the 'preoccupied' condition.

However, its number of senses in WordNet (2) is one unit lower than for ‘concerned’ (3).

We can observe that the notion of ‘worrying’ highly dominates when rephrasing ‘concerned’, while ‘thinking’ is the word most involved when rephrasing the alternative wording, ‘preoccupied’. It thus seems that these two alternatives are not fully equivalent. Moreover, only 16% of participants in the ‘preoccupied’ condition used the word ‘concern’ to define it, and only one of those in the ‘concerned’ condition used the word ‘preoccupied’. There were also several definitions that were exclusive to one of the two conditions. On the one hand, ‘anxiety’, ‘afraid’ and ‘scared’ are associated with ‘concerned’. On the other hand, ‘busy’, ‘on mind’, ‘affected’, ‘focused heavily’, ‘informed’ and ‘involved’ were only used to define ‘preoccupied’.

In addition, we compared the two samples. Those in the general sample were more likely to use the wording ‘worried’ and some definitions were only used by them, i.e., ‘preoccupied’, ‘anxious’ and ‘affected’. On the other hand, non-native speakers were more likely to use the wording ‘concerned’, and the word ‘afraid’ was used only by participants in this group.

5.2.2 Summary and discussion

In summarising the above findings, we observe the results of the cognitive interviews with respect to the level of similarity when the alternative wordings were used.

A relatively high match (i.e., both alternatives share large portions of similar response categories) was observed for the following cases:

- **‘**Careful’ and ‘cautious’ (P0);
- ‘Threat’ and ‘menace’ (P1);
- ‘Ransom money’ and ‘demanded money’ (P5).

In these three cases, respondents were presented with the lower frequency wording using its high-frequency alternative to define or paraphrase it; for example, ‘careful’ was used to define the word ‘cautious’, ‘threat’ was often used in paraphrases of ‘menace’, and ‘ransom’ was used in paraphrases of ‘demanded money’. In the high-frequency conditions, on the other hand, the low-frequency alternatives were either less

commonly used (‘cautious’ was rarely used to define ‘careful’) or not used at all (‘menace’ was not used to paraphrase ‘threat’, and ‘demanded money’ was not used to paraphrase ‘ransom money’).

A somewhat lower level (i.e., the alternatives share similar response categories but to a smaller extent than in the above, highly matched cases) of similarity can be attributed in the following cases:

- ‘Sympathetic to’ and ‘support’ (P4.2 and P4.3);
- ‘Collecting’ and ‘gathering’ (P4.5);
- ‘Prone to’ and ‘inclined to’ (P6);
- ‘Chances’ and ‘risk’ (P4.1).

Among the abovementioned cases, only ‘sympathetic’ and ‘inclined’ were paraphrased with their more frequent alternatives – that is, ‘support’ and ‘prone’, respectively; while the opposite was found for the high-frequency word ‘gathering’, which was more often paraphrased with the low-frequency alternative ‘collecting’ than the reverse. ‘Chances’ and ‘risk’, on the other hand, were never used to paraphrase each other.

In the remaining cases, there is a very low level of match as they do not share similar response categories. Thus, we might argue that the following alternatives have relatively different meanings and are not real synonyms:

- ‘Worried’ and ‘apprehensive’ (P2);
- ‘Justified’ and ‘vindicated’ (P3);
- ‘Restricting’ and ‘limiting’ (P4.4);
- ‘Encourage’ and ‘promote’ (P7);
- ‘Concerned’ and ‘preoccupied’ (P8).

Even when there was a low level of match, the high-frequency alternative was sometimes used to define the low-frequency alternative. For instance, ‘worried’ was used to define ‘apprehensive’, ‘justified’ was used to define ‘vindicated’, and ‘concerned’ was used to define ‘preoccupied’, but not the reverse. On the other hand, ‘restricting’ and ‘limiting’ were only rarely used to paraphrase each other; the same held true for ‘encourage’ and ‘promote’.

The above differences will help us in the final interpretation, where they will be compared with corpora frequencies and integrated with the interpretations from expert interviews and outcomes from the quantitative study. We can also highlight two additional observations:

- There was sometimes more variation in participant's answers in the case of some high-frequency wordings, indicating that they might have less clear and more ambiguous meanings. In most cases, this was reflected in the different number of senses in WordNet (e.g., 'cautious' and 'careful'), while in some it was not (e.g., 'apprehensive' and 'worried').
- Several differences between native and non-native speakers were observed; for instance, 'cautious' was less often used by non-native speakers than by those in the general sample. In addition, some definitions and paraphrases were used exclusively by non-native speakers, while others were used only by native speakers. However, due to sample size, it is not easy to generalise the corresponding conclusions.

6 Main split-ballot experiment

In this chapter, we present the main study, which was based on a split-ballot experiment using a text corpora approach to evaluate question wordings. However, as opposed to the study described in Chapter 3, where students were used, this study was conducted on a general population. In addition, four versions of the same questionnaire were used (and not just two). Moreover, instead of focusing solely on single-word frequencies, we also evaluated the entire wording frequencies for the related strings of words. For example, instead of examining one single word (e.g., 'threat'), the specific context in which this word appeared was also analysed, which included the key neighbouring words (e.g., 'threat of terrorism'). This enabled us to operate with a much more real and specific context of the use of a certain word. The frequencies were thus calculated for corresponding strings of words, not just for single words. The main advantage of this study, however, is the fact that we use the same cases that were analysed in the studies presented in Chapter 4 and 5, where we conducted the selection of the question items, as well as extensive expert evaluations and cognitive interviews for the majority of question items and wording cases, so that we could compare the corpora approaches with expert evaluations and cognitive interviews also in the light of response quality.

The case study addresses the selection of PEW Internet research questions on the sensitive topic of terrorism, which were already evaluated with text corpora frequencies and expert reviews in Chapter 4. Due to specifics of this survey experiment, we also analysed some additional questions and wordings. Since we used the same approach and methodology (and from the same source) elaborated in Chapters 3 and 4, we do not need to repeat the details here; however, we do need to include some material that is essential to the flow and understanding of this experiment.

Based on various alternative wordings, we formed four versions of the questionnaire with different levels of wording frequencies. In total, 16 question items were subject to variation of 38 cases. We should mention here that by 'case' we mean an example of a word (or a set of words) which is then subject to replacement with alternative words (synonyms). Most of these cases typically have only one alternative, so we have two options per case: the original and the alternative wordings. Sometimes, however, we also have two alternatives (and three wordings). In total, we thus have 81 wording alternatives allocated across the four versions of the questionnaire. Survey results were

then compared according to a series of survey data quality indicators, which is also related to the main research question: How do the text corpora frequencies of alternative wordings affect data quality indicators?

We first present how we selected the wording alternatives for the experiment. We also describe the empirical setting, the Survey Monkey Audience panel (https://www.surveymonkey.com), which was used to collect the data. Next, we analyse the results, starting with socio-demographic characteristics and response distributions, followed by the analysis of selected response quality indicators. Finally, we summarise the results.

6.1 Methodology

We conducted an experiment based on a questionnaire constructed from a selection of PEW research questions on the sensitive topic of terrorism (a broader introduction to PEW was already elaborated in Chapter 4). Based on queries on linguistic corpora and lexical databases, we developed four versions of the questionnaire with different levels of difficulty, where certain wordings were replaced with synonymous terms with different wording frequencies:

- The initial version retained the original format from PEW research (labelled as '0')
- The second version replaced 12 wordings with more frequent wordings (improved version, labelled as '1')
- The third version replaced original wordings with less frequent wordings: It made 16 replacements (worse version, labelled as '-1')
- The fourth version further deteriorated Version -1 by using much less frequent word alternatives: It made 18 more replacements than Version '-1', for a total of 34 replacements (the worst version, labelled as '-2').

6.1.1 Selection of items, cases and alternative wordings

The questions were selected from a repository of PEW questions, as fully elaborated in Chapter 4 (Section 4.1.2). Table 6.1 is basically a replication of Table 4.2; however, it

has additional items, denoted with an asterisk (*), which were introduced (as further elaborated below) to reinforce the effects of the variations in word frequencies.

The first column denotes the item number and is shaded for the line, which presents the original **case** – that is, the word for which we sought alternatives with similar meanings (synonyms). The second column ('Insert') repeats and describes the relation to the alternative word according to the search in the WordNet database: The options are hyponym, hypernym, see also, similar to, and so forth. The third column ('Synset') reports the alternative synonym rings – so-called 'synsets' – for the case in question. In the fourth column, the synset definitions from WordNet are presented.

Table 6.1: Table of initial alternative wording (synonyms) based on WordNet database

Case	Query	Synset	Definition
*P0	cautious	**careful** (adj)	exercising caution or showing care or attention
	see also	**cautious** (adj)	showing careful forethought
	see also	diligent (adj)	characterized by care and perseverance in carrying out tasks
	see also	prudent (adj)	careful and sensible; marked by sound judgment
	similar to	certain, sure (adj)	exercising or taking care great enough to bring assurance
	similar to	… [7 other similar words] (adj)	… [various definitions]
P1	threat	**menace, threat** (n)	something that is a source of danger
	hyponym	yellow peril (n)	the threat to Western civilization said to arise from the power of Asiatic peoples
	hypernym	**danger** (n)	a cause of pain or injury or loss
*P1.II	reducing	repress, quash, keep down, **subdue,** subjugate, **reduce** (v)	put down by force or intimidation
	hypernym	oppress, suppress, crush (v)	come down or keep down by unjust use of one's authority
P2	worried	**apprehensive, worried** (adj)	mentally upset over a possible misfortune or danger, etc.
	similar to	**uneasy** (adj)	lacking sense of security or affording no ease or reassurance
	other meaning	disquieted, distressed, disturbed, **upset, worried** (adj)	afflicted with or marked by anxious uneasiness or trouble or grief
	similar to	troubled (adj)	characterized by or indicative of distressed or affliction or danger or need
	other PoS	**concern**, interest, occupy, worry (v)	be on the mind of
*P2.II	attack	**attack,** onslaught, onset, onrush (n)	(military) an offensive against an enemy (using weapons)
	hyponym	… [12 hyponyms]	… [various definitions]
	meronym	assault (n)	close fighting during the culmination of a military attack
	hypernym	operation, military operation (n)	activity by a military or naval force (as a

Case	Query	Synset	Definition
			maneuver or campaign)
	other meaning	**attack, attempt** (n)	the act of attacking
P3	justified	**justify, vindicate** (v)	show to be right by providing justification or proof
	troponym	**excuse**, explain (v)	serve as a reason or cause or justification of
	troponym	**legitimate** (v)	show or affirm to be just and legitimate
	troponym	**warrant** (v)	provide adequate grounds to justify (a certain course of action)
	hypernym	uphold, maintain (v)	support against an opponent
*P3.II	suspected	**suspected** (adj)	believed likely
	MS word thesaurus	supposed, alleged, so-called, assumed (adj)	doubtful or suspect
	MS word thesaurus	**hypothetical**, theoretical, imaginary, mad-up, fictional, invented (adj)	based primarily on surmise rather than adequate evidence
*P3.III	gain	**acquire,** win, **gain** (v)	win something through one's efforts
	troponym	cozen (v)	cheat or trick
	hypernym	get, **acquire** (v)	come into the possession of something concrete or abstract
*P4.0	part	**part,** portion, component part, component, **constituent** (n)	something determined in relation to something that includes it
	constitute (as verb form of constituent)	**constitute**, represent, make up, comprise, be (v)	form or compose
P4.1	chances	**probability, chance** (n)	a measure of how likely it is that some event will occur, a number expressing the ratio of favorable cases to the whole number of cases possible
	hyponym	**risk**, risk of exposure (n)	the probability of being exposed to an infectious agent
	hyponym	… [7 other hyponyms] (n)	… [various definitions]
	hypernym	measure, quantity, amount (n)	how much there is or how many there are of something that you can quantify
*P4.1.I	worry	**worry** (v)	be worried, concerned, anxious, troubled, or uneasy
	definition (anxious)	**anxious**, nervous, queasy, uneasy, unquiet (adj)	causing or fraught with or showing anxiety
P4.2 P4.3	sympathetic to	**sympathetic** (adj)	expressing or feeling or resulting from sympathy or compassion or friendly fellow feelings, disposed towards
	definition	**disposed to**	naturally disposed toward
	see also	**compassionate** (adj)	showing or having compassion
	see also	congenial (adj)	suitable to your needs
	see also	**kind** (adj)	having or showing a tender or considerate and helpful nature
	similar to	commiserative (adj)	feeling or expressing sympathy
	similar to	condolent (adj)	expressing sympathy with a person who experienced the death of a loved one
	similar to	empathic, empathetic (adj)	showing empathy or ready comprehension
	other PoS	feel for, pity, compassionate, condole	share the suffering of

Case	Query	Synset	Definition
		with, **sympathize with** (v)	
	MS Word thesaurus	**favor**, favour (n)	an inclination to approve
	MS Word thesaurus	**supportive** (adj)	furnishing support or assistance
	other PoS	**support**, back up (v)	give moral or psychological support, aid, or courage to
*P4.2.II	extend	widen, **broaden, extend** (v)	extend in scope or range or area
	troponym	… [4 troponyms]	… [various definitions]
	hypernym	increase (v)	make bigger or more
*P4.3.II	allowed	let, **allow**, **permit** (v)	make it possible through a specific action or lack of action for something to happen
	troponym	pass (v)	allow to go without comment or censure
*P4.3. III	search	research, **search**, explore (v)	inquire into
	troponym	… [5 troponyms]	… [various definitions]
	hypernym	**investigate**, look into (v)	investigate scientifically
*P4.3. IV	court	**court, tribunal**, judicature (n)	an assembly (including one more judges) to conduct judicial business
	hyponym	… [26 hyponyms]	… [various definitions]
	hypernym	assembly	a group of persons who are gathered together for a common purpose
P4.4	restricting	**restrict, curtail, curb, cut back (v)**	place restrictions on
	troponym	**abridge** (v)	lessen, diminish, or curtail
	troponym	immobilize, immobilise (v)	cause to be unable tomove
	troponym	ration (v)	restrict the consumption of a relatively scarce commodity, as during war
	troponym	restrict, **control** (v)	place under restrictions; limit access to by law
	hypernym	**limit**, circumscribe, confine to (v)	restrict or confine within limits
*P4.4.II	too	excessively, overly, to a fault, too (adv)	to a degree exceeding normal or proper limits
P4.5	collecting	**gather, garner, collect, pull together** (v)	assemble or get together
	definition	**assembling**	collect in one place
	troponym	… [19 other troponyms]	… [various definitions]
*P4.5.II	like	**like, similar** (adj)	resembling or similar, having the same or some of the same characteristics; often used in combination
	see also	same (adj)	closely similar or comparable in kind or quality or quantity of degree
	similar to	… [3 other similar words]	… [various definitions]
P4.6.1	overwhelming	**overpowering, overwhelming** (adj)	so strong as to be irresistible
	similar to	irresistible resistless (adj)	impossible to resist, overpowering
*P4.6.II	way	**manner**, mode, style, **way**, fashion (n)	how something is done or how it happens
	hyponym	… [9 hyponyms]	… [various definitions]
	hypernym	property	a basic or essential attribute shared by all members of a class
*P4.6.2	defeat	get the better of, **overcome, defeat**	win a victory over

Case	Query	Synset	Definition
		(v)	
	troponym	demolish, destroy (v)	defeat soundly and humiliatingly
	troponym	beat, beat out, crush, shell, trounce, vanquish	come out better in a competition, race, or conflict
	troponym	… [19 other troponyms]	… [various definitions]
	other meaning	kill, shoot down, **defeat**, vote down, vote out (v)	thwart the passage of
P5a	ransom money	**ransom, ransom money**	money demanded for the return of a captured person
	definition	**demanded** (v)	request urgently and forcefully
	hypernym	cost (n)	the total spent for goods or services including money and time and labor
P5b	hostages	**hostage, surety** (n)	a prisoner who is held by one party to insure that another party will meet specified terms
	hypernym	prisoner, captive (n)	a person who is confine; especially a prisoner of war
P6	prone	**prone** (adj)	having a tendency (to)
	similar to	**inclined** (adj)	having a preference, disposition, or tendency
*P6.II	view	position, view, perspective (n)	a way of regarding situations or topics etc.
	hyponym	… [8 other hyponyms]	… [various definitions]
	hypernym	orientation (n)	an integrated set of attitudes and beliefs
P7	encourage	**promote, advance, boost, further, encourage** (v)	contribute to the progress or growth of
	troponym	… [7 other troponyms]	… [various definitions]
	hypernym	support, back up (v)	give moral or psychological support, aid, or courage to
*P7.III	believers	**believer, worshiper**, worshipper (n)	a person who has religious faith
	hyponym	… [9 hyponyms]	… [various definitions]
	hypernym	religious person (n)	a person who manifests devotion to a deity
P8	concerned	**concerned** (adj)	feeling or showing worry or solicitude
	see also	attentive (adj)	giving care or attention
	see also	**troubled** (adj)	characterized by or indicative of distressed or affliction or danger or need
	similar to	afraid (adj)	filled with regret or concern; used often to soften an unpleasant statement
	similar to	**afraid** (adj)	filling worry or concern or insecurity
	similar to	haunted, obsessed, **preoccupied**, taken up (adj)	having or showing excessive or compulsive concern with something
	similar to	**solicitous** (adj)	full of anxiety and concern
	other PoS	**concern**, interest, occupy, **worry** (v)	be on the mind of

For all bolded wordings in Table 6.1, we calculated the wording frequencies in Table 6.2. We may add here that in the expert evaluation research presented in Chapter 4, in

order to limit the response burden of the experts, we only selected one case per question, which was then varied with alternative wordings. Typically, it was a noun, but it could have also been a verb, adjective or adverb in the question. We should also repeat that when there was more than one relevant case in one question item, the selection criteria in expert evaluation was the frequency level of the original wording as well as how many alternatives were available. For instance, for item P1 ('In general, how well do you think the US government is doing in reducing the threat of terrorism?'), we preferred to choose 'threat' for the expert evaluation instead of 'reducing', as the latter had a lower frequency (in enTenTen) and more wording alternatives (two instead of only one).

However, as mentioned above, in the split-ballot experiment, in addition to the cases selected for expert evaluation, we also included other cases wherever there was an opportunity for a case with good alternatives (synonyms). For example, in item P1, we also included the case 'reducing' (which was explicitly rejected as the best case for expert evaluation), as illustrated in Table 6.2.

Table 6.2 below provides wording frequencies for alternative wordings. A higher frequency basically reflects a higher appearance of a certain word in text corpora (the text corpora that we used, enTenTen and COCA, were introduced and described in Chapter 4). Similar to the previous table, for the majority of cases, the frequencies of synonyms and other alternative wordings were already presented in Chapter 4. Here too, the wordings that were not yet elaborated with expert reviews in Chapter 4 are denoted with an asterisk (*). In fact, as mentioned above, some of the wordings listed in the table were not used in expert reviews (to limit the response burden of the experts), but were used here so we could take full advantage of the possibility to modify the wordings with synonyms. Correspondingly, wherever there was a realistic opportunity in a certain item to replace a certain word with an alternative, we checked for it and, if suitable, created an additional case with alternative wordings.

Table 6.2: Complete table of corpora frequencies based on COCA and Ten-ten

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
*P0	careful	23408	833568	careful in dealing	5	147
	cautious	5617	180225	cautious in dealing	7	85
P1	danger	20370	512584	danger of terrorism	8	116
	menace	1866	50138	menace of terrorism	2	126
	threat	30382	666925	threat of terrorism	192	2209
*P1.II	reducing	14008	782021	reducing the threat	31	602
	subduing	159	4778	subduing the threat	0	2
P2	apprehensive	920	18993	how apprehensive	4	30
	concerned	39502	776023	how concerned	114	963
	uneasy	3386	35863	how uneasy	6	104
	upset	15417	265608	how upset	122	1696
	worried	25024	324153	how worried	181	824
*P2.II	attack	51754	3008661	terrorist attack	1471	65935
	attempt	33421	2804183	terrorist attempt	2	404
P3	excused	1272	17263	ever excused	0	8
	legitimate	10844	279690	ever legitimate	3	33
	justified	5038	111269	ever justified	10	227
	vindicated	742	8863	ever vindicated	1	3
	warranted	1595	36895	ever warranted	2	31
*P3.II	torture	7126	351772	use of torture against suspected	0	14
	torturing	655	21926	torturing suspected	1	42
*P3.III	suspected	11981	272533	suspected terrorists	306	3749
	hypothetical	2753	66564	hypothetical terrorists	0	5
*P3.IV	gain	25656	3091890	gain information	91	5891
	acquire	7020	1613498	acquire information	62	3994
*P4.0	occasional	10617	174986	occasional acts	4	111
	periodic	2946	125855	periodic acts	1	13
*P4.0.II	part	259169	13662023	be part of life	5	878
	constitute	6097	433976	constitute life	3	102
P4.1	chances	12915	482356	chances of attack	1	49
	probability	5075	222141	probability of attack	2	28
	risk	64294	2526058	risk of attack	15	483
*P4.1II	worry	35696	2080814	often worry	28	2415
	anxious	8951	271186	often anxious	16	378
P4.2-3	compassionate to	40	1022	compassionate to terrorists	0	0
	disposed to	352	7501	disposed to terrorists	0	0
	in favour of	213	65112	in favour of terrorists	0	1
	kind to	1251	38483	kind to terrorists	0	2
	support	120828	6192586	support terrorists	18	458
	supportive of	1948	29212	supportive of terrorists	1	3

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
	sympathetic to	1319	12456	sympathetic to terrorists	1	18
	sympathize with	712	13030	sympathize with terrorists	0	16
*P4.2II	extend	11914	1470710	extend to groups	0	39
	broaden	1909	140758	broaden to groups	0	0
*P4.3II	allowed	52767	1939253	allowed to search	4	377
	permitted	7651	370033	permitted to search	0	69
*P4.3II	search	41014	5607506	search the houses	2	161
	investigate	10967	839351	investigate the houses	1	7
*P4.3IV	court	123076	4223826	court order	1251	70690
	tribunal	2539	135449	tribunal order	0	480
P4.4	abridging	56	906	abridging liberties	0	1
	controlling	4	2179	controlling liberties	0	0
	curbing	610	12729	curbing liberties	0	1
	curtailing	319	4859	curtailing liberties	0	6
	cutting back	1000	15578	cutting back liberties	0	0
	limiting	4514	116859	limiting liberties	1	4
	restricting	1665	39914	restricting liberties	1	7
*P4.1II	too	368019	12471272	too far	8232	217221
	excessively	1216	73728	excessively far	0	30
P4.5	assembling	1481	32420	assembling information	7	83
	collecting	7573	208996	collecting information	154	4417
	garnering	308	10618	garnering information	1	37
	gathering	11133	340008	gathering information	312	7542
	pulling together	199	4261	pulling together information	0	48
*P4.1II	like	1064398	35727539	people like me	904	26798
	similar	72355	3331139	people similar to me	1	66
*P4.6a	overwhelming	10092	358582	overwhelming force	144	2573
	overpowering	921	31884	overpowering force	10	194
*P4.6b	defeat	11725	703885	defeat terrorism	38	774
	overcome	11956	720464	overcome terrorism	1	52
*P4.6II	way	545164	24642664	way to defeat	52	2806
	manner	21500	1399695	manner to defeat	1	7
P5a	demanded money	41	856	demanded money for hostages	0	0
	ransom	1330	27098	ransom for hostages	5	9
	ransom money	41	506	ransom money for hostages	0	1
P5b	hostages	4627	15729	money for hostages	1	1
	sureties	7	1570	money for surities	0	0
P6	inclined	4716	118399	inclined to violence	3	39
	prone	3884	140703	prone to violence	59	452
*P6.II	is	4823632	248380312	is closer	858	29237

Single word (grey shade used for the original wording)		Freq COCA	Freq enTenTen	String of words	Freq COCA	Freq enTenTen
	comes	108759	5138057	comes closer	196	4681
*P6.III	views	28551	1383554	your own views	11	1105
	perspective	29806	1284950	your own perspective	4	1042
P7	advance	17699	1438599	advance violence	0	4
	boost	9625	603199	boost violence	0	2
	encourage	17136	1311056	encourage violence	9	517
	further	64650	2472985	further violence	69	922
	promote	15942	1488081	promote violence	29	1002
*P7.II	believers	3036	231200	among its believers	0	50
	worshippers	489	22985	among its worshippers	0	16
P8	afraid	31099	411099	afraid about extremism	0	0
	concerned	39502	776023	concerned about extremism	1	3
	preoccupied	2148	21312	preoccupied about extremism	0	0
	solicitous	335	2327	solicitous about extremism	0	0
	troubled	8576	111411	troubled about extremism	0	0
	worried	25024	324153	worried about extremism	0	0
*P9	consider	55697	7636685	consider yourself	382	19113
	reckon	1434	173879	reckon yourself	0	85

6.1.2 Identification of cases with changes in wording across four versions

Following the above elaboration, the words were correspondingly alternated in four versions of the questionnaire. Let us first observe the slight improvements (Version 1) and slight deterioration (Version -1) of the original wording (Table 6.3). The cases (words or sequences of words) that were changed in each item are in light-grey shading (Table 6.3). When there were two wordings changed within a single item (the items P3, P4.6, P5, P6 and P7), the second wording was shaded in dark grey to differentiate it from the first case. As mentioned, the second alteration was introduced to increase the potential differences.

Table 6.3 presents how the original version changed into a positive 'Version 1' (Improved) or a negative 'Version -1' (Worse), while Error! **Not a valid bookmark self-reference.** shows further changes of 'Version -1' which were made to attain

'Version -2'. For example, the case 'careful' (frequency: 147) was replaced with the word 'cautious' (85).

Table 6.4 presents how the worse version was further worsened in 'Version -2' (Worst).

Let us look, for instance, at the word 'concerned', which is more frequent (963) in the specific context than the alternative, 'worried' (824) (see Table 6.2); so, this may introduce slight improvements in 'Version 1', while expert evaluation endorsed both options equally. We may also add here that the option 'upset' was eliminated due to unfavourable expert evaluations: Despite the fact that it was identified as a synonym based on a WordNet search, its meaning is actually rather different. As an extreme option for 'Version -1', the word 'apprehensive' (with a frequency of only 30) was selected.

Table 6.3: Cases in first three versions (0, 1 and -1)

ORIGINAL VERSION (Version 0)	**WORSE (Version -1)**	**IMPROVED (Version 1)**
P0. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	-	-
P1. In general, how well do you think the U. S. government is doing in reducing the threat of terrorism?	menace	-
P2. How worried are you that there will soon be another terrorist attack in the United states?	apprehensive	concerned
P3. Do you think the use of torture against suspected terrorists in order to gain important information can ever be justified?	vindicated	torturing;
P4. To what extent do you agree or disagree with the following statements? Please select an answer for each statement.	-	-
P4.0. Occasional acts of terrorism in the U.S. will be part of life in the future.	periodic	-
P4.1. I often worry about the chances of a nuclear attack by terrorists.	probability	risk
P4.2. Freedom of speech should not extend to groups that are sympathetic to terrorists.	disposed towards	support
P4.3. The police should be allowed to search the houses of people who might be sympathetic to terrorists without a court order.	disposed towards	support
P4.4. The government anti-terrorism policies have gone too far in restricting the average person's civil liberties.	curtailing	limiting
P4.5. I am concerned that the government is collecting too much information about people like me.	assembling	gathering
P4.6. Using overwhelming military force is the best way to defeat	overpowering	overwhelming;

ORIGINAL VERSION (Version 0)	WORSE (Version -1)	IMPROVED (Version 1)
terrorism.		overcome
P5. As you may know, the United States government has a policy that it NEVER pays ransom money for hostages held by terrorist groups. What is your opinion about this policy?	sureties	demanded money
P6. Which statement comes closer to your own views even if neither is exactly right? Some religions are more prone to violence than others. All religions are about the same when it comes to violence.	inclined	is; -
P7. Which statement comes closer to your own views even if neither is exactly right? The Islamic religion is more likely to encourage violence among its believers. The Islamic religion does not encourage violence more than other.	boost	is; promote
P8. How concerned, if at all, are you about Islamic extremism around the world these days?	preoccupied	-
P9. In politics today, do you consider yourself a Republican, Democrat, or independent?	reckon	-

Error! Not a valid bookmark self-reference. shows further changes of 'Version -1' which were made to attain 'Version -2'. For example, the case 'careful' (frequency: 147) was replaced with the word 'cautious' (85).

Table 6.4 Comparison of changes between Version -1 and Version -2

WORSE Version -1	WORST Version -2
P0. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	cautious
P1. In general, how well do you think the U. S. government is doing in reducing the menace of terrorism?	subduing
P2. How apprehensive are you that there will soon be another terrorist attack in the United states?	attempt
P3. Do you think the use of torture against suspected terrorists in order to gain important information can ever be vindicated?	hypothetical; acquire
P4. To what extent do you agree or disagree with the following statements? Please select an answer for each statement.	-
P4.0. Periodic acts of terrorism in the U.S. will be part of life in the future.	constitute
P4.1. I often worry about the probability of a nuclear attack by terrorists.	am often anxious; attempt
P4.2. Freedom of speech should not extend to groups that are disposed towards terrorists.	be broadened

WORSE Version -1	WORST Version -2
P4.3. The police should not be allowed to search the houses of people who might be disposed towards terrorists without a court order.	permitted; investigate; tribunal
P4.4. The government anti-terrorism policies have gone too far in curtailing the average person's civil liberties.	excessively
P4.5. I am concerned that the government is assembling too much information about people like me.	similar to me
P4.6. Using overpowering military force is the best way to defeat terrorism.	manner
P5. As you may know, the United States government has a policy that it NEVER pays ransom money for sureties held by terrorist groups. Overall, do you approve or disapprove of this policy?	ransom
P.6 Which statement comes closer to your own views even if neither is exactly right? Some religions are more inclined to violence than others. All religions are about the same when it comes to violence.	perspective
P7. Which statement comes closer to your own views even if neither is exactly right? The Islamic religion is more likely to boost violence among its believers. The Islamic religion does not boost violence more than others.	perspective; worshippers

6.1.3 Overview of all wording changes

We can summarise all cases into an overview of alternative wordings for all four versions (Table 6.5). In total, as previously mentioned, within 16 items we alternated 38 cases with 81 different wordings. The majority of cases had only two alternatives; however, in items P2, P4.2, P4.3, P4.5 and P5, there were cases with three alternative wordings.

Table 6.5: Overview of items, cases and wording included into four versions of the questionnaire

Item wording	Version -2	Version -1	Version	Version 1
P0. Cautious/careful in dealing	Cautious	Careful	Careful	Careful
P1. Subduing/reducing the menace/threat of terrorism	Subduing Menace	Reducing Menace	Reducing Threat	Reducing Threat
P2. How apprehensive/worried … terrorist attempt/attack	Apprehensive Attempt	Apprehensive Attack	Worried Attack	Concerned Attack

Item wording	Version -2	Version -1	Version	Version 1
P3. Use of torture/torturing … hypothetical/suspected … acquire/gain information … ever vindicated/justified	Use of torture Hypothetical Acquire Vindicated	Use of torture Suspected Gain Vindicated	Use of torture Suspected Gain Justified	Torturing Suspected Gain Justified
P4.0. Occasional/periodic acts … will constitute/be part of life	Periodic Constitute	Periodic Be part of	Occasional Be part of	Occasional Be part of
P4.1 (am) often anxious/worry … probability/chances/risk of attempt/attack	Anxious Probability Attempt	Worry Probability Attack	Worry Chances Attack	Worry Risk Attack
P4.2 Be broadened/extend to groups … disposed towards/ sympathetic to/support terrorists	Be broadened Disposed tow.	Extend Disposed tow.	Extend Sympathetic	Extend Support
P4.3 Permitted/allowed to investigate/search … disposed towards/sympathetic to/support terrorists … tribunal/court	Permitted Investigate Disposed tow. Tribunal	Allowed Search Disposed tow. Court	Allowed Search Sympathetic Court	Allowed Search Support Court
P4.4 Excessively/too far in curtailing/restricting/limiting liberties	Excessively Curtailing	Too Curtailing	Too Restricting	Too Limiting
P4.5 Assembling/collecting/ gathering information … similar to/like me	Assembling Similar to	Assembling Like	Collecting Like	Gathering Like
P4.6 Overpowering/ overwhelming force … manner/way …overcome/defeat	Overpowering Manner Defeat	Overpowering Way Defeat	Overwhelming Way Defeat	Overwhelming Way Overcome
P5. Ransom/demanded (money) for sureties/hostages	Ransom Sureties	Ransom m. Sureties	Ransom m. Hostages	Demanded m. Hostages
P6. Comes/is closer … perspective/views … inclined/prone to violence	Comes Inclined Perspective	Comes Inclined Views	Comes Prone Views	Is Prone Views
P7. Comes/is closer … perspective/views … boost/ encourage/promote violence … worshippers/believers	Comes Perspective Boost Worshippers	Comes Views Boost Believers	Comes Views Encourage Believers	Is Views Promote Believers
P8. Preoccupied/concerned about extremism	Preoccupied	Preoccupied	Concerned	Concerned
P9. Reckon/consider yourself	Reckon	Reckon	Consider	Consider

6.1.4 The experimental design

The study was carried out on the Survey Monkey Audience, a non-probability online panel recruited from a diverse population of more than 30 million unique individuals who visit the Survey Monkey website (to create or take surveys) every month (Survey

Monkey 2015). Those who sign up to become members are invited to complete a detailed profile that includes targeting criteria such as gender, age, household income, education, race, etc. In addition, benchmarking surveys are regularly carried out to make sure members are representative of the US adult population. Apart from email invitations to panel members, there are also so-called 'routed responses' where respondents are recruited from other active surveys, wherein they are occasionally presented with an invitation to take an additional survey. When a potentially cooperative respondent agrees to participate, he or she is randomly routed to an active survey deployment.

According to the official information on the Survey Monkey website (Survey Monkey 2015), the members are awarded for their participation with non-cash incentives, such as charitable donations and sweepstakes entries. This motivates respondents and, in contrast to monetary incentives, does not encourage unwanted response behaviours. In addition, there is a limit on the number of surveys a respondent can take per week in order to prevent them from participating too many times.

We ordered a random sample of 2,400 units (US residents, 18 years or older) from Survey Monkey, 600 for each of the four experimental groups (i.e., Versions), to which the respondents were randomly allocated. However, to guarantee the desired number of responses was reached, more than 2,400 invites were sent out, and due to the randomness of selection, an unequal number of respondents started each survey.

The project was run as a stand-alone questionnaire on 1 October and 2 October, 2015. The data were collected in less than 24 hours. The title 'Extremism Concerns' was displayed to respondents.

6.2 Results

In this section, we first present the demographic characteristics and response distributions of the questions included in the experiment. Next, we systematically evaluate the four versions of the questionnaire with various aspects of survey data quality: drop-out, response times, 'don't know' response, acquiescence, and subjective evaluations. Our default hypothesis throughout this dissertation is that in versions where alternative wordings (mostly synonyms but also other words or strings of words with

similar meanings) with higher text corpora frequency were used, the survey data produced was of a higher quality.

6.2.1 Socio-demographic structure of the sample

In total, there were 2,966 invited persons who started responding (we do not have the information on the total number of invites sent): 780 to the worst version (Version -2), 719 to the worse version (Version -1), 739 to the version with the original wordings (Version 0), and 730 to the version with the improved wordings (Version 1). Although there were more respondents to the worst version than to the other three, the difference in the number of starting units was not statistically significant (Chi-square = 1.41; *df* = 3; $p = 0.70$).

Due to drop-out and item nonresponse, the number of units who actually responded to each question was lower and varied from question to question. Besides the five background variables provided by Survey Monkey (gender, age, household income, US region and device type) that were included in the database, we also asked respondents about their race, religion, education level, native language and self-assessed ability to read a book or newspaper in English. The responses to all 10 socio-demographic variables, broken down by questionnaire version, are presented in Table 6.5, including the number of respondents for each item.

Table 6.6: Socio-demographic characteristics of the four sample versions complete responses

Variable	-2 Worst	-1 Worse	0 Original	1 Improved
Race/ethnicity	**671**	**653**	**668**	**654**
American Indian or Alaskan Native	1.5%	3.2%	1.0%	1.4%
Asian / Pacific Islander	4.5%	6.3%	3.9%	5.4%
Black or African American	6.0%	4.6%	4.6%	6.0%
Hispanic American	5.2%	6.1%	5.4%	7.5%
White / Caucasian	82.9%	79.8%	85.0%	79.8%
Multiple ethnicity / Other (please specify)	0.0%	0.0%	0.1%	0.0%
Religion (multiple responses possible)	**646**	**633**	**639**	**630**
Protestant	26.6%	24.1%	26.3%	24.2%
Roman Catholic	21.3%	20.7%	20.8%	20.0%
Mormon	1.5%	1.5%	1.4%	3.2%
Orthodox (Greek or Russian)	1.1%	1.2%	1.7%	1.8%
Jewish	4.5%	3.4%	4.1%	2.9%

Variable	-2 Worst	-1 Worse	0 Original	1 Improved
Muslim	1.4%	1.1%	0.6%	0.9%
Buddhist	2.4%	3.1%	1.8%	2.6%
Hindu	1.4%	1.1%	0.9%	0.9%
atheist	9.2%	10.2%	10.1%	11.7%
agnostic	10.1%	11.4%	9.8%	8.5%
nothing in particular	27.6%	28.9%	28.7%	30.4%
Other	11.2%	10.3%	11.5%	12.2%
Education	**661**	**646**	**653**	**647**
Less than high school degree	1.7%	1.5%	2.1%	1.7%
High school degree of equivalent	8.6%	10.4%	11.2%	10.0%
Some college but no degree	21.2%	20.6%	18.8%	21.8%
Associate's degree	10.9%	10.5%	8.7%	9.0%
Bachelor's degree	31.8%	28.9%	31.5%	30.3%
Graduate degree	25.9%	28.0%	27.6%	27.2%
Native language	**661**	**645**	**653**	**646**
English	94.6%	93.2%	92.5%	91.0%
Spanish	2.1%	1.9%	2.0%	2.9%
Chinese	0.5%	0.3%	0.3%	1.2%
Tagalog	0.9%	0.8%	0.5%	0.9%
French	0.3%	0.6%	0.2%	0.6%
Vietnamese	0.0%	0.3%	0.3%	0.5%
German	0.3%	0.3%	0.8%	0.2%
Korean	0.2%	0.5%	0.0%	0.2%
Other	1.2%	2.2%	3.5%	2.5%
Self-assessed ability to read a newspaper/book in English	**661**	**645**	**652**	**645**
Very well	94.9%	94.1%	93.4%	93.2%
Pretty well	3.3%	4.2%	4.6%	5.0%
Just a little	1.2%	0.8%	1.2%	1.1%
Not at all	0.6%	0.9%	0.8%	0.8%
Gender	**646**	**633**	**640**	**630**
Female	51.1%	50.4%	50.3%	50.2%
Male	48.9%	49.6%	49.7%	49.8%
Age	**646**	**633**	**640**	**630**
18-30	24.8%	24.0%	23.8%	23.7%
30-44	24.1%	24.8%	24.5%	25.9%
45-59	25.7%	25.0%	25.5%	24.6%
> 60	25.4%	25.2%	26.3%	29.9%
Household income	**646**	**633**	**639**	**630**
$0 to $9,999	5.0%	7.1%	7.7%	4.9%
$10,000 to $24,999	7.0%	8.5%	8.1%	9.0%
$25,000 to $49,999	16.6%	17.2%	13.5%	13.8%
$50,000 to $74,999	18.3%	18.3%	17.5%	18.6%
$75,000 to $99,999	16.6%	12.6%	18.0%	16.5%
$100,000 to $124,999	12.1%	10.1%	7.8%	9.5%
$125,000 to $149,999	5.0%	4.7%	6.1%	5.9%
$150,000 to $174,999	2.5%	2.7%	3.0%	3.0%
$175,000 to $199,999	1.7%	3.5%	3.6%	1.7%
$200,000 and up	5.9%	5.5%	4.9%	6.3%
Prefer not to answer	9.6%	9.6%	9.9%	10.6%
US region	**646**	**633**	**639**	**630**

Variable	-2 Worst	-1 Worse	0 Original	1 Improved
New England	5.0%	6.1%	6.4%	6.9%
Middle Atlantic	14.0%	13.2%	14.2%	13.4%
East North Central	16.8%	18.9%	14.2%	13.7%
West North Central	6.6%	4.8%	6.5%	7.7%
South Atlantic	14.8%	15.2%	18.1%	16.6%
East South Central	5.5%	5.2%	6.1%	4.2%
West South Central	9.9%	7.7%	10.1%	10.1%
Mountain	6.8%	8.4%	5.6%	8.3%
Pacific	20.6%	20.5%	18.8%	19.1%
Device type	**646**	**633**	**640**	**630**
iOS Phone / Tablet	28.3%	29.5%	28.4%	26.3%
Android Phone / Tablet	18.6%	15.8%	17.7%	19.2%
Other Phone / Tablet	0.0%	0.2%	0.0%	0.0%
Win. Desktop / Laptop	43.8%	44.1%	43.4%	43.7%
MacOS Desktop / Laptop	7.4%	9.0%	9.1%	9.2%
Other	1.9%	1.4%	1.4%	1.6%

In general, we can conclude that the sample reflects the US adult population: about half of the sample was female and about half male. There was also about a quarter of respondents from each of the four age groups. Respondents belonged to various household income groups and resided in different US regions. About 55% responded on desktop computers or laptops, while the remaining 45% responded using mobile devices.

Most of the sample was of Caucasian origin, ranging from 80% to 85% in different groups. Almost half of the respondents selected at least one religion when asked about their current religion. Almost 60% of the sample had at least a bachelor's degree. English was the native language for more than 90% of the sample, and between 93% and 95% responded that they were able to read a newspaper or book in English very well.

We tested if the distributions in the four groups (versions) significantly differed from the overall distribution. The Pearson chi-square statistic was used to evaluate if there were any statistically significant differences in the response distributions, and Cramer's V was computed to evaluate the effect sizes (Table 6.7). To avoid small cells, religion, language and self-assessed ability to read were recoded into only two categories before conducting the test.

Table 6.7: Pearson chi-square and Cramer's V for differences in demographic structure across versions

	Pearson chi-square			Cramer's V	
	Value	df	Sig. (2-sided)	Value	Approx Sig.
Gender	0.1	3	1.00	0.1	1.00
Age	2.9	12	1.00	0.02	1.00
Income	37.1	30	0.18	0.07	0.18
Region	25.6	24	0.38	0.06	0.38
Device	8.9	15	0.89	0.03	0.89
Race	**22.1**	**12**	**0.04**	**0.05**	**0.04**
Religion (selected at least one)	0.72	3	0.87	0.02	0.87
Education	9.6	16	0.90	0.06	0.90
Language (is English)	6.3	3	0.10	0.05	0.10
Read (answered Very well)	2.0	3	0.58	0.03	0.58

***Bolded rows** are those where the chi-square statistic is statistically significant ($p < 0.05$).

The only statistically significant difference between the four groups was for race. It appears that the sample of respondents in Version 0 had a larger share of the 'white' race. However, the race question was asked towards the end of the survey, so it was confounded with differential drop-out rates between the four groups. Thus, we cannot determine if the sample was biased or if it was an effect of a higher drop-out of a certain demographic segment. In any case, the difference is quite small and the corresponding effects for our research are negligible.

6.2.2 Response distributions across the four versions of the questionnaire

Sixteen question items were subject to wording changes in our experiment. Since all of the alternatives were synonymous wordings and respondents were randomly assigned to one of the four versions, we did not expect any differences in their response distributions, as presented in Table 6.8.

Table 6.8: Response distributions for question items with single response format

Variable	-2 Worst	-1 Worse	0 Original	1 Improved
P0. Generally speaking, would you say that most people …	**780**	**719**	**737**	**730**
Generally speaking, most people can be trusted	45.1%	46.3%	42.5%	46.3%
You can't be too careful in dealing with people	48.8%	46.9%	51.3%	47.0%
Don't know	6.0%	6.8%	6.2%	6.7%
P1. In general, how well do you think the United States …	**771**	**713**	**732**	**718**
Very well	7.3%	5.9%	4.9%	7.4%
Fairly well	35.5%	35.9%	39.8%	39.0%
Not too well	29.4%	31.8%	28.1%	25.8%
Not at all well	19.5%	17.1%	19.4%	17.5%
Don't know	8.3%	9.3%	7.8%	10.3%
P2. How worried are you that there will soon be another …	**763**	**709**	**727**	**710**
Very worried	16.1%	13.1%	16.9%	21.7%
Somewhat worried	44.4%	40.9%	45.7%	39.2%
Not too worried	23.5%	30.0%	25.6%	26.1%
Not at all worried	7.9%	10.3%	7.8%	8.3%
Don't know	8.1%	5.6%	4.0%	4.8%
P3. Do you think the use of torture against suspected …	**757**	**705**	**722**	**707**
Often justified	12.9%	12.1%	16.9%	14.9%
Sometimes justified	26.3%	26.8%	30.6%	25.7%
Rarely justified	21.1%	23.8%	23.4%	23.2%
Never justified	24.8%	24.4%	20.5%	27.0%
Don't know	14.8%	12.9%	8.6%	9.2%
P4.0 Occasional acts of terrorism in the U.S. will …	**608**	**621**	**624**	**606**
Completely agree	20.4%	25.3%	28.8%	26.7%
Mostly agree	54.1%	52.8%	54.3%	47.9%
Mostly disagree	20.2%	16.9%	13.5%	18.6%
Completely disagree	5.3%	5.0%	3.4%	6.8%
P4.1 I often worry about the chances of a nuclear attack …	**653**	**642**	**639**	**617**
Completely agree	8.1%	9.7%	12.5%	11.3%
Mostly agree	28.0%	29.3%	30.2%	26.9%
Mostly disagree	39.7%	40.7%	39.3%	37.6%
Completely disagree	24.2%	20.4%	18.0%	24.1%
P4.2 Freedom of speech should not extend to groups that …	**617**	**620**	**624**	**620**
Completely agree	18.3%	18.4%	17.6%	21.3%
Mostly agree	25.4%	24.0%	20.8%	24.1%
Mostly disagree	29.7%	31.6%	30.8%	28.8%
Completely disagree	26.6%	26.0%	30.8%	25.9%
P4.3 The police should be allowed to search the ...	**640**	**635**	**645**	**620**
Completely agree	15.3%	15.3%	12.4%	13.7%
Mostly agree	24.1%	19.2%	16.1%	17.4%

Variable	-2 Worst	-1 Worse	0 Original	1 Improved
Mostly disagree	28.6%	32.1%	31.3%	29.7%
Completely disagree	32.0%	33.4%	40.2%	39.2%
P4.4 The government's anti-terrorism policies have gone …	**616**	**605**	**611**	**576**
Completely agree	15.3%	18.2%	13.9%	16.8%
Mostly agree	32.0%	29.1%	28.2%	26.0%
Mostly disagree	35.2%	36.9%	40.4%	37.3%
Completely disagree	17.5%	15.9%	17.5%	19.8%
P4.5 I am concerned that the government is collecting too …	**628**	**624**	**627**	**605**
Completely agree	18.8%	20.5%	22.0%	23.3%
Mostly agree	26.4%	29.2%	29.0%	26.1%
Mostly disagree	32.8%	32.7%	33.8%	28.9%
Completely disagree	22.0%	17.6%	15.2%	21.7%
P4.6 Using overwhelming military force is the best …	**620**	**601**	**598**	**599**
Completely agree	18.7%	20.0%	25.8%	16.7%
Mostly agree	27.9%	28.0%	22.9%	25.4%
Mostly disagree	27.3%	28.3%	26.1%	31.1%
Completely disagree	26.1%	23.8%	25.3%	26.9%
P5. As you may know, the United States government ...	**685**	**669**	**680**	**663**
Approve	56.8%	62.5%	56.0%	61.7%
Disapprove	14.0%	12.1%	18.4%	14.8%
Don't know	29.2%	25.4%	25.6%	23.5%
P6. Which statement comes closer to your own views …	**683**	**667**	**677**	**660**
Some religions are more prone to violence than others	48.3%	51.9%	56.1%	50.5%
All religions are about the same when it comes to violence	36.7%	33.1%	30.1%	33.2%
Don't know	14.9%	15.0%	13.7%	16.4%
P7. Which statement comes closer to your own views …	**680**	**663**	**675**	**659**
The Islamic religion is more likely to encourage violence among its believers	40.4%	39.2%	39.7%	38.7%
The Islamic religion does not encourage violence more than others	36.5%	34.8%	36.7%	37.2%
Don't know	14.9%	15.0%	13.7%	16.4%
P8. Which statement comes closer to your own views …	**678**	**657**	**674**	**657**
Very concerned	7.2%	5.3%	38.1%	33.8%
Somewhat concerned	29.1%	29.1%	39.2%	40.9%
Not too concerned	37.5%	39.1%	12.6%	15.2%
Not at all concerned	18.6%	19.5%	3.7%	2.6%
Don't know	7.7%	7.0%	6.4%	7.5%
P9. In politics today, do you consider yourself ...	**677**	**656**	**674**	**655**
Republican	22.9%	21.3%	23.6%	21.2%
Democrat	36.9%	34.8%	33.7%	34.5%
Independent	33.8%	36.9%	35.2%	36.6%
Other	6.4%	7.0%	7.6%	7.6%

The Pearson chi-square test was computed to evaluate if the response distributions across the four groups differed from the joint distribution. As a result, we actually did find certain statistically significant differences, denoted in bold, for some items, as shown in **Error! Not a valid bookmark self-reference.**.

Table 6.9: Chi-Square test and Cramer's V for differences in response distributions across versions

	Pearson chi-square			Cramer's V	
	Value	df	Sig. (2-sided)	Value	Approx Sig.
P0. Cautious/careful in dealing with people	4.0	6	0.67	0.04	0.67
P1. Subduing/reducing the menace/threat of terrorism	17.1	12	0.15	0.07	0.15
P2. How apprehensive/worried … terrorist attempt/attack	**42.9**	**12**	**0.00**	**0.12**	**0.00**
P3. Use of torture/torturing … hypothetical/suspected … acquire/gain information … ever vindicated/justified	**35.9**	**12**	**0.00**	**0.11**	**0.00**
P4.0. Occasional/periodic acts … will constitute/be part of life	**28.6**	**9**	**0.00**	**0.11**	**0.00**
P4.1 (am) often anxious/worry … probability/chances/risk of attempt/attack	10.2	9	0.33	0.06	0.33
P4.2 Be broadened/extend to groups … disposed towards/sympathetic to/support	17.2	9	0.05	0.08	0.05
P4.3 Permitted/allowed to investigate/search … disposed towards/sympathetic to/support terrorists … tribunal/court	**25.1**	**9**	**0.00**	**0.10**	**0.00**
P4.4 Excessively/too far in curtailing/restricting/limiting liberties	12.6	9	0.18	0.07	0.18
P4.5 Assembling/collecting/gathering information … similar to/like me	**18.4**	**9**	**0.03**	**0.09**	**0.03**
P4.6 Overpowering/overwhelming force … manner/way …overcome/defeat	**21.6**	**9**	**0.01**	**0.09**	**0.01**
P5. Ransom/demanded (money) for sureties/hostages	**17.5**	**6**	**0.01**	**0.08**	**0.01**
P6. Comes/is closer … perspective/views … inclined/prone to violence	**10.3**	**6**	**0.11**	**0.06**	**0.11**
P7. Comes/is closer … perspective/views … boost/encourage/promote violence … worshippers/believers	2.1	6	0.91	0.03	0.91
P8. Preoccupied/concerned about extremism	**611.1**	**12**	**0.00**	**0.48**	**0.00**

***Bolded rows** are those where the chi-square statistic is statistically significant ($p < 0.05$).

There are statistically significant differences for items P2, P3, P4.1, P4.3, P4.5, P4.6, P5, P6 and P8. For instance, when using 'apprehensive' instead of 'worried', respondents were more likely to respond that they were not worried. When using 'justified' and not 'legitimate', there were more of those who agreed with the statement. Using 'overpowering' instead of 'overwhelming' decreased agreement with the statement on military power. 'Demanded' instead of 'ransom' increased agreement with the statement on US policy. Replacing 'prone' with 'inclined' increased the share of those who believed that some regions' groups are more inclined towards violence. Lastly, the largest differences were found when substituting 'concerned' with 'preoccupied', which decreased the reported level of concern.

These differences might also indicate that the wording alternatives are not real synonyms, as they affect not only response quality but also response distributions. Of course, these effects could also be the result of a different sample structure across versions. However, there were few differences in socio-demographic variables among the four versions, so it is somewhat more likely that these effects are coming from substantive differences in wordings (i.e., the wordings are not true synonyms). Still, we should not underestimate the potentially differential impact of the self-selection process (related to the drop-outs) on sample structure in each version. Yet, in most cases, these effects are relatively small – the Cramer's V value ranges from 0.06 to 0.12 – so this should not interfere with our general comparisons of response quality indicators for the four versions of the questionnaire. Nevertheless, the specific differences in certain cases can help us integrate the interpretations from corpora frequencies, experts, cognitive interviews and the survey experiment.

On the extreme side, for item P8, the differences between the versions are substantial and the effect size is large. We would not have expected such a difference in response distributions for two items that are supposedly synonymous, but apparently 'concerned' and 'preoccupied' do not convey the same meaning and the latter is perhaps too extreme. Thus, response quality indicators for this item should be interpreted with caution.

6.2.3 Drop-outs

As mentioned, out of the 2,966 persons who started responding, 2,557 (86%) reached the last page. However, the corresponding drop-out rate (14% overall) varied across the four versions: the original, improved and worse versions had a drop-out rate of about 12-13%, while the worst version had a drop-out rate of about 17%. According to the ANOVA test, the difference is statistically significant ($F = 3.2$; $df = 3$; $p = 0.02$) but the effect is relatively small (Cohen's d = 0.19).

In addition, we also checked where drop-outs occurred and found a peak on Page 5 (Table 6.10), where the only matrix question was located. This is not surprising, as this response format is often found to be difficult. Moreover, there are again differences in response distributions between versions ($F = 4.6$; $df = 3$; $p = 0.00$), but the effect size is small (Cohen's d = 0.24).

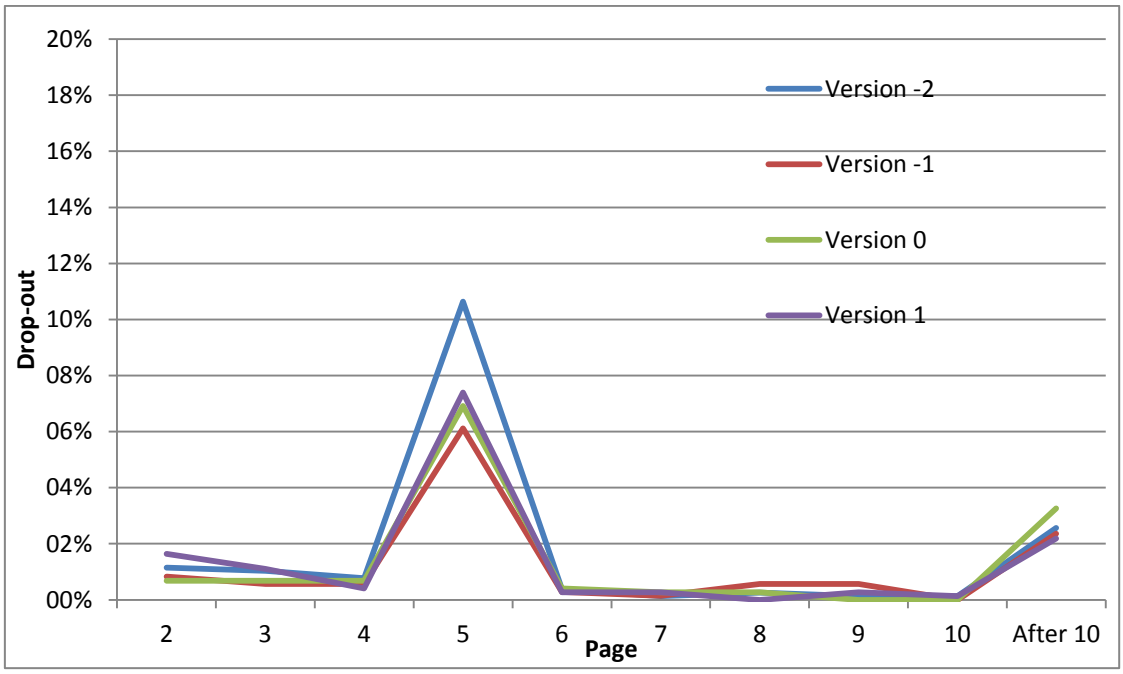
Table 6.10: Drop-out location across pages and items

Pages and items	Version -2	Version -1	Version 0	Version 1
Page 2 (item P1)	1.2%	0.8%	0.7%	1.6%
Page 3 (item P2)	1.0%	0.6%	0.7%	1.1%
Page 4 (item P3)	0.8%	0.6%	0.7%	0.4%
Page 5 (items P4.0-P4.6)	10.6%	5.8%	6.9%	7.3%
Page 6 (item P5)	0.4%	0.3%	0.4%	0.3%
Page 7 (item P6)	0.1%	0.1%	0.3%	0.3%
Page 8 (item P7)	0.3%	0.6%	0.3%	-
Page 9 (item P8)	0.1%	0.6%	-	0.3%
Page 10 (item P9)	0.1%	-	-	0.1%
Page 11- 21 (items P10+)*	2.6%	2.4%	3.1%	1.9%
Total drop-out	132 17%	84 12%	96 13%	97 13%

* Items P10-P16 relate to variables, which were the same for all versions

Figure 6.1 illustrates the effect of the matrix question, while the other items share the drop-out rate relatively homogeneously.

Figure 6.1: Drop-out location



Based on these results, we can claim that the respondents who received the worst wordings were more prone to dropping out of the survey. This is in line with our expectations – we presumed that higher drop-out rates were a consequence of the relatively bigger burden that the respondents in Version -2 were exposed to.

6.2.4 Response times

We analysed survey response times as an indicator of response quality based on the assumption that a higher average time may reflect higher cognitive effort. Unfortunately, Survey Monkey could not provide the time stamps for each page, so the analysis is limited only to total survey response times. Thus, detailed latency measures such as those proposed by Fazio (1990) could not be computed.

The median and average response times were computed for all four versions. However, both statistics are skewed due to drop-outs and outliers, especially averages, so we also computed the median and average times for the subset of those who completed the survey (i.e., reached the last page) as well as those who did so in less than about 13 minutes, as 95% of respondents completed the survey in 783 seconds or less (Table 6.11). We may add that there were no differences in this percentage (which was computed and cut at the level of the entire sample) across the four versions. Similarly,

there were no differences across the versions in the share of speeders – that is, the fastest 5% of the units.

Table 6.11: Median and average response times access the four versions

	-2 Worst	-1 Worse	0 Original	1 Improved
Completes	**648**	**635**	**641**	**633**
Average time	12m 37s	8m 18s	7m 15s	5m 48s
Median time	4m 53s	4m 50s	4m 35s	4m 37s
Complete in 783s or less (about 13m)	**613 95%**	**600 95%**	**611 95%**	**605 96%**
Average time	5m 11s	5m 3s	4m 56s	4m 56s
Levene's test for equality of variances;	F = 1.1, p = 0.28	F=0.03, p=0.87	-	F=0.17 p=0.68
Independent samples t-test for difference from V0	t = 2 , p = 0.05 d = 0.11, r = 0.08	t = 0.995, p = 0.3 d = 0.06, r = 0.04		t = -0.02, p = 0.98 d = 0.00, r = 0.00
Median time	4m 44s	4m 43s	4m 29s	4m 31s
Mann-Whitney test for difference from V0	w=203290 p = 0.01	w = 197960 p = 0.11	-	w = 181310 p = 0.57

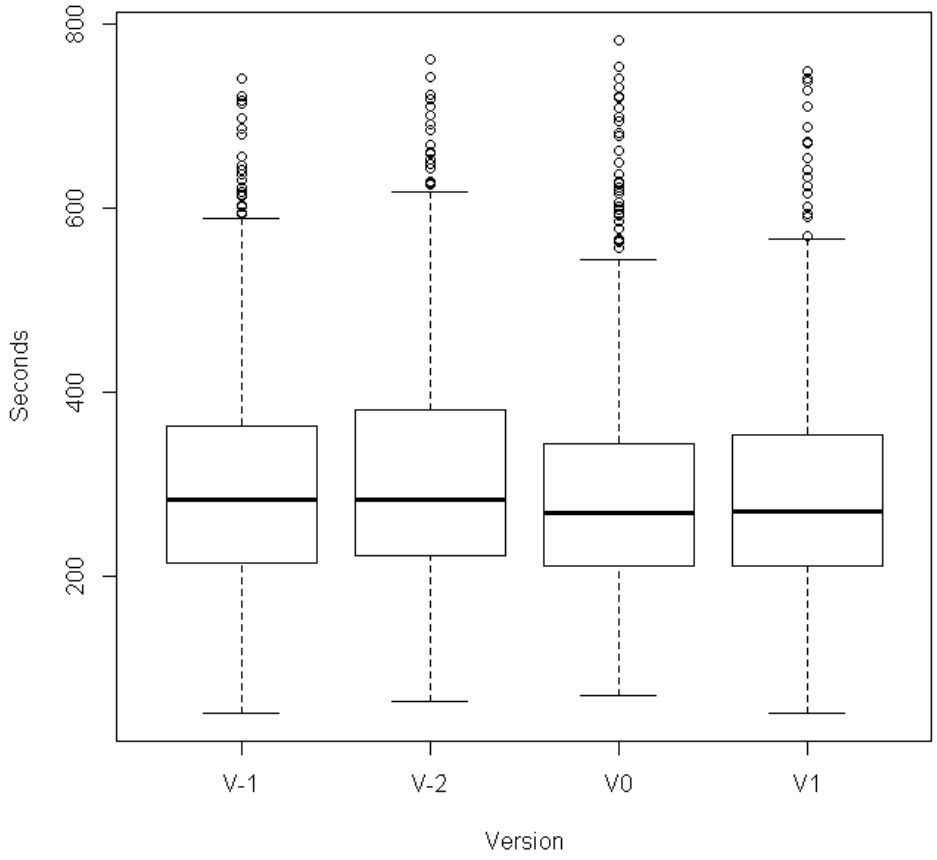
On average, respondents to the worst (-2) version were the slowest, followed by the worse (-1) group, and then the original (0) and improved (1) groups, which had the quickest respondents. This is also the expected order, as it reflects the order of the alternative wordings based on corpora frequencies.

However, looking only at responses completed in less than 13 minutes, the differences are much smaller, and specifically there is no difference between the original (0) and improved (1) versions. Moreover, when analysing medians, which are a more conservative measure, the differences are generally even smaller, especially when looking only at those who completed the survey in less than 13 minutes, where the original version had a lower response time than the improved one.

The differences were also tested with different statistical tests. We ran the independent sample *t*-test for the difference between Version 0 and the other three versions, which showed that there is only a significant difference between the original and the worst versions ($p = 0.05$). For median times, we used the non-parametric Kruskal-Wallis test, which showed statistically significant differences between the four groups ($H = 7.9$; $df = 3$; $p = 0.047$). In addition, we conducted the Mann-Whitney U test for pairs of groups.

The worst version significantly differs from the original version. On the other hand, there are no significant differences between the original and worse versions, nor between the original and improved versions, as can be observed in Figure 6.2.

Figure 6.2: Time variability in different groups



To summarise, the analysis of response times leads to similar findings as the analysis of drop-outs. On average, respondents allocated to Version -2 needed more time to complete a survey – presumably due to the increased cognitive burden. However, there are no significant differences among the other three versions, although there are certain indications (however, not throughout consistently) that versions with the wordings based on higher corpora frequencies provided slightly shorter response times. First, the 5% of outliers (which were cut off in the analysis) seem to spend much less time in versions with higher corpora frequencies. Moreover, even after cutting off the outliers, the order of the total response times is in line with expectations based on the supposed difficulty of the versions.

6.2.5 'Don't know' responses

Except for item P9, all other items that were subject to changes in wording also included a 'don't know' (DK) option, which we presumed would be more often selected in cases where questions were difficult for respondents to understand. For each of them, we computed the share of DK responses, excluding units with incomplete responses (drop-outs). We also computed the median DK percentage. The results of the computation for the 15 items with the DK option are presented in Table 6.12.

Table 6.12: Don't know responses across the four versions of the questionnaire

	Version -2	Version -1	Version 0	Version 1
P0. Cautious/careful in dealing with people	5.7%	6.8%	5.8%	6.5%
P1. Subduing/reducing the menace/threat of terrorism	7.3%	8.3%	6.7%	9.0%
P2. How apprehensive/worried … terrorist attempt/attack	6.0%	5.2%	3.4%	4.7%
P3. Use of torture/torturing … hypothetical/suspected … acquire/gain information … ever vindicated/justified	11.3%	11.0%	7.6%	8.2%
P4.0. Occasional/periodic acts … will constitute/be part of life	11.1%	7.4%	8.3%	8.7%
P4.1 (am) often anxious/worry … probability/chances/risk of attempt/attack	4.8%	4.4%	6.4%	7.1%
P4.2 Be broadened/extend to groups … disposed towards/sympathetic to/support	10.2%	7.7%	8.3%	8.1%
P4.3 Permitted/allowed to investigate/search … disposed towards/sympathetic to/support terrorists … tribunal/court	6.9%	5.4%	5.1%	6.5%
P4.4 Excessively/too far in curtailing/restricting/limiting liberties	10.5%	9.8%	10.6%	13.0%
P4.5 Assembling/collecting/gathering information … similar to/like me	8.8%	7.1%	7.8%	8.7%
P4.6 Overpowering/overwhelming force … manner/way …overcome/defeat	9.6%	10.4%	12.6%	9.6%
P5. Ransom/demanded (money) for sureties/hostages	28.2%	25.0%	25.4%	23.1%
P6. Comes/is closer … perspective/views … inclined/prone to violence	14.2%	14.5%	13.4%	15.6%
P7. Comes/is closer … perspective/views … boost/encourage/promote violence … worshippers/believers	22.7%	25.4%	22.6%	23.4%
P8. Preoccupied/concerned about extremism	7.3%	7.1%	6.1%	7.3%
Median DK %	9.6	7.7	7.8	8.7

The median percentage of the worst group is higher than that of the other three groups; however, the Mann-Whitney U test showed that there are no significant differences between either pair of medians. Furthermore, we observed differences in percentages for individual items. The first two items, P0 and P1 (and also item P7), have more DK answers in the worse and improved version, while the original and the worst version have fewer of them. For items P2, P3 and P5, on the other hand, there is a higher share of DK answers in Version -2 than in the other three versions, followed by Version -1. Next, for most items in the matrix question as well as for item P8, the highest share of DK answers is also in Version -2 (in P4.0, P4.2, P4.2 and P4.5), but it is then followed by Version 1, not Version -1. Items P4.1 and P4.4 in the matrix (and also item P6) have the most DK answers in the improved version. The only item where the original version has the highest share of DK answers is item P4.6 in the matrix question.

However, the only significant difference in DK answers is for item P3 (Chi-square = 7.8; df = 3; p = 0.05), where most of the DK answers are in the worst and worse versions (11%), while the share of DK responses in the improved and original versions is three percentage points lower (8%). Moreover, the effect size is very small (Cramer's V is 0.06).

In addition, we computed the total number of DK answers across all 15 items, as shown in Table 6.13.

Table 6.13: Number of DK responses across all items

		-2 Worst	-1 Worse	0 Original	1 Improved
Counts of DK responses		**649**	**635**	**641**	**633**
	0 items	42.0%	40.5%	44.8%	44.2%
	1 item	24.5%	27.9%	23.1%	25.0%
	2 items	12.3%	12.6%	13.3%	12.5%
	3 items	6.5%	6.8%	7.8%	6.2%
	4 items	4.5%	4.9%	4.4%	3.0%
	5-9 items	7.7%	4.9%	4.4%	5.8%
	10+ items	2.5%	2.5%	2.3%	3.3%

About 40 to 45% of respondents provided a substantial answer to all 15 items, while other respondents selected the DK option at least once. The distribution of DK responses is about the same across all four groups, and no significant differences could be observed. In addition, we also computed the Spearman correlation rank between the number of DK responses and the versions assuming the variable has an ordinal scale (-2, -1, 0, 1), as the versions can be ordered from the version with wordings with the lowest corpora frequencies to the version with the highest corpora frequencies. However, the correlation is only -0.03 and thus not statistically significant.

We can conclude that except for one item, there are no significant differences in the share of DK answers between the four versions. Thus, we cannot demonstrate a relationship between the wording changes and the tendency to select the DK option; except maybe for item P3, where we replaced 'justified' with the much less frequent wording 'vindicated' in the worse and worst version. In addition, in the worst version, the word 'suspected' was also replaced with the less frequent 'hypothetical', as well as 'gain' with the less frequent 'acquire'. However, since there is no difference between Version -1 and Version -2, the effect of the wording changes between Version 0 and -1 (and, similarly, Version 0 and -2) should only be attributable to the 'justified–vindicated' replacement.

6.2.6 Acquiescence

Another indicator of response quality we evaluated is acquiescence, the tendency to agree with statements. It is assumed that when an item is more difficult to comprehend, respondents tend to agree more. We could only compute it for the matrix question (items P4.0–6), for which we had an agreement scale.

Specifically, we decided to compute the sum of the 'agreed' and 'completely agreed' response for each statement. Thus, we obtained seven variables, where value 1 means 'agreed' or 'completely agreed', while the two 'disagree' options were recoded to 0. First, we analysed them separately, item by item. Table 6.14 shows how many respondents agreed or completely agreed with the statements in the matrix question.

Table 6.14: Acquiescence levels across the four versions of the questionnaire

% Completely Agree + Agree	Version -2	Version -1	Version 0	Version 1
P4.0. Occasional/periodic acts … will constitute/be part of life	66.0%	72.4%	75.7%	68.9%
P4.1 (am) often anxious/worry … probability/chances/risk of attempt/attack	34.1%	38.0%	39.3%	35.7%
P4.2 Be broadened/extend to groups … disposed towards/sympathetic to/support terrorists	39.7%	39.2%	35.6%	42.5%
P4.3 Permitted/allowed to investigate/search … disposed towards/sympathetic to/support terrorists … tribunal/court	37.0%	32.4%	26.7%	29.5%
P4.4 Excessively/too far in curtailing/restricting/limiting liberties	42.1%	42.8%	37.3%	37.6%
P4.5 Assembling/collecting/gathering information … similar to/like me	40.7%	46.6%	46.8%	45.5%
P4.6 Overpowering/overwhelming force … manner/way …overcome/defeat	42.6%	43.1%	42.3%	37.9%
Median acquiescence %	**40.1**	**42.8**	**39.3**	**37.9**

The median acquiescence for the worst (40.1) and particularly for the worse (42.8) conditions is higher than for the original (39.3) and improved (37.9) conditions. This is in line with the assumption that there will be more acquiescence in the case of worse or worst wordings; however, the Mann-Whitney U-test showed no significant differences between either pair of medians.

Moreover, we observed differences in percentages for individual items. Only for items P4.3, P.4.4 and P4.6 is there more agreement in the worse and worst version. On the other hand, items P4.0, P4.1 and P4.5 have more agreement in the case of the original wording. Similarly, for item P4.2, there is more agreement in the case of the improved version; however, the worst and worse options have more acquiescence than the original version.

Next, we ran an ANOVA to test if the differences were statistically significant and found that they are only significant for items P4.0 and P4.3. The effect size is small, though, as we can see in Table 6.15.

Table 6.15: ANOVA for acquiescence across the four versions of the questionnaire

	Mean Square		F test		Effect size (Cohen's d)
	Between Groups	Within Groups	F	Sig.	
P4.0. Occasional/periodic acts … will constitute/be part of life	**1.1**	**0.2**	**5.5**	**0.00**	**0.29**
P4.1 (am) often anxious/worry … probability/chances/risk of attempt/attack	0.2	0.2	1.5	0.22	-
P4.2 Be broadened/extend to groups … disposed towards/sympathetic to/support terrorists	0.5	0.2	2.2	0.09	0..19
P4.3 Permitted/allowed to investigate/search … disposed towards/sympathetic to/support terrorists … tribunal/court	**1.3**	**0.2**	**5.9**	**0.00**	**0.30**
P4.4 Excessively/too far in curtailing/restricting/limiting liberties	.05	0.2	2.3	0.08	0.15
P4.5 Assembling/collecting/gathering information … similar to/like me	0.5	0.2	2.1	0.10	
P4.6 Overpowering/overwhelming force … manner/way …overcome/defeat	0.4	0.2	1.5	0.21	-

*__Bolded__ rows are those where the F statistic is statistically significant ($p < 0.06$).

There was more of a tendency to agree with statement P4.0 in Version 0 as compared to other versions. It thus seems that replacing the wording 'occasional' with the less frequent 'periodic' in P4.0 did not have the predicted effect (but there is a tendency towards the opposite effect). For item P4.3, on the other hand, the finding goes in the predicted direction: replacing 'sympathetic to' with the less frequent 'disposed towards', and 'court' with the less frequent 'tribunal', might be a factor that increases the tendency to acquiesce. However, the difference in the share of those who agreed could also be attributed to a shift in question meaning if the changed wordings are not really synonymous.

A better measure of acquiescence might be the percentage of those who selected 'completely agree' or 'agree' across a series of items, as those who tend to acquiesce usually tend to do it for several items. Table 6.13 shows how many respondents selected one of these two options in zero to seven items.

Table 6.16: Acquiescence across all seven items

	-2 Worst	-1 Worse	0 Original	1 Improved
Count of completely agree + agree	**649**	**635**	**641**	**633**
0 items	9.0%	5.5%	7.0%	8.1%
1 item	12.2%	12.6%	12.9%	12.3%
2 items	19.3%	19.8%	20.9%	21.0%
3 items	21.0%	24.1%	22.5%	23.9%
4 items	18.7%	15.3%	15.0%	14.2%
5 items	10.2%	12.1%	12.2%	11.8%
6 items	5.1%	5.5%	5.9%	5.2%
7 items	4.6%	5.0%	3.6%	3.5%

More than 90% of the respondents agreed or completely agreed with at least one item, but there were only a few who agreed or completely agreed with all seven items. The percentage of those respondents was lower in the original and improved versions (about 3.5%) compared to those in the worse and worst versions (between 4.6 and 5%). However, the difference is not statistically significant (Chi-square = 2.8; df = 3; p = 0.42).

Nevertheless, the above percentages do not show the full picture – it should be also taken into account that two of the items in the matrix question, P4.4 and P4.5, represent opinions that are somewhat opposite to those expressed by other items. This is also confirmed by their low correlation with other acquiescence variables as measured by Spearman correlation ranks. Thus, we computed a factor score (using the principal axis factoring method) and computed its Spearman correlation with the versions as an ordinal variable (-2, -1, 0, 1). However, the correlation is only 0.02 and thus not significant.

We can argue that variation in wording has relatively weak and inconsistent effects on acquiescence. Thus, the default hypothesis that synonyms with higher word frequencies may decrease the acquiescence level cannot be confirmed.

6.2.7 Subjective evaluations of respondents

Finally, we analysed the subjective evaluations of respondents. There were three questions at the end of the survey that measured how much respondents enjoyed completing the questionnaire (S1), how difficult it was for them to interpret the meaning of questions (S2), and how many words they found at least a little difficult to understand (S3). We assumed that respondents to the more difficult (worse and worst) versions enjoyed responding less, found it more difficult to interpret meanings of questions, and reported a higher number of words they could not understand.

In Table 6.17, we present the responses to the first two questions and their average values, broken down by questionnaire version.

Table 6.17: Subjective evaluations across the four versions

	Version -2	Version -1	Version 0	Version 1
S1. How much did you enjoy completing the questionnaire?	**656**	**642**	**649**	**640**
1 - Not at all well	13.6%	13.2%	12.6%	12.5%
2 - A little	22.6%	18.7%	16.2%	20.3%
3 - A moderate amount	35.4%	41.1%	39.8%	38.3%
4 - A lot	16.5%	15.1%	18.2%	16.9%
5 - A great deal	12.0%	11.8%	13.3%	12.0%
Average	**4.6**	**4.6**	**4.7**	**4.7**
S2. How difficult was it for you to interpret the meanings of questions in this questionnaire?	**655**	**642**	**649**	**640**
1 - Extremely difficult	1.1%	1.2%	0.8%	0.8%
2 - Very difficult	2.4%	1.1%	1.1%	1.1%
3- Moderately difficult	8.1%	6.5%	6.6%	5.9%
4 - Slightly difficult	15.3%	15.3%	9.7%	10.8%
5 - Not difficult at all	73.1%	75.9%	81.8%	81.4%
Average	**4.8**	**3.9**	**3.0**	**4.6**

It appears that those who responded to the original version enjoyed it the most, while respondents who responded to the worse or worst version enjoyed it much less. Moreover, they also found it more difficult to interpret the meaning of the questions.

However, to confirm this relation, we ran some statistical tests. Table 6.18 presents the results of the Pearson chi-square across the four versions.

Table 6.18: Pearson chi-square and Cramer's V for subjective evaluation questions across versions

	Pearson chi-square			Cramer's V	
	Value	df	Sig. (2-sided)	Value	Approx Sig.
S1. How much did you enjoy completing the questionnaire?	12.1	12	0.43	0.07	0.43
S2. How difficult was it for you to interpret the meanings of questions in this questionnaire?	**27.1**	**12**	**0.01**	**0.06**	**0.01**

There were no statistically significant differences in terms of enjoyment, but differences do exist for difficulty interpreting the meaning of questions. We can illustrate this with the order for the share of the 'Not difficult at all' category – Version -2 (73.1%), Version -1 (75.9%) and Version 0 (81.1%) – which actually follows our initial assumption about the potential positive impact of high corpora frequencies on survey data quality.

In addition, we also ran ANOVA for both variables, the results of which are presented in Table 6.19.

Table 6.19: ANOVA for subjective evaluation questions across the four versions of the questionnaire

	Mean Square		F test		Effect size (Cohen's d)
	Between Groups	Within Groups	F	Sig.	
S1. How much did you enjoy completing the questionnaire?	1.3	1.4	1.00	0.39	-
S2. How difficult was it for you to interpret the meanings of questions in this questionnaire?	**3.0**	**.06**	**5.47**	**0.00**	**2.4**

Again, the result is not statistically significant for enjoyment but is significant for difficulty; further, the effect size is big (Cohen's d = 2.4).

Response difficulty is affected not only by questionnaire characteristics but also by participant characteristics. Thus, we also checked the interaction between variable S2 and three other variables: education, native language, and gender. As expected, men found the questionnaire to be more difficult than women (Chi-square = 13.7; *df* = 4; *p* = 0.01), non-native speakers found it to be more difficult than native speakers (Chi-square = 64.7; *df* = 4; *p* < 0.01), and those who were less educated found it to be more difficult than those who were more educated (Chi-square = 43.2; *df* = 4; *p* < 0.01). After controlling for the relationship between questionnaire difficulty and the version variable (-2, -1, 0, 1) with language, we could confirm differences between versions for native speakers (Chi-square = 28.1; *df* = 12; *p* = 0.01) but not for non-native speakers (Chi-square = 8.3; *df* = 12; *p* = 0.77). Similarly, the association is still significant for those who were highly educated (Chi-square = 27.8; *df* = 12; *p* = 0.01) but not for those who were not (Chi-square = 7.1; *df* = 12; *p* = 0.85). When controlling for gender, the association between difficulty and version is still present, but it is very weak, both for males (Chi-square = 19.7; *df* = 12; *p* = 0.07) and females (Chi-square = 18.6; *df* = 12; *p* = 0.10).

Table 6.20 reports how many words respondents found to be at least a little difficult to understand for each version.

Table 6.20: Number of words that were at least a little difficult to understand across the four versions

S3. When you were reading the questions in this survey, about how many words in the questions were at least a little difficult to understand?	Version -2	Version -1	Version 0	Version 1
	649	**635**	**641**	**633**
10+ words	2.5%	2.4%	0.9%	1.6%
5-9 words	4.5%	2.5%	1.7%	2.7%
4 words	2.8%	2.0%	1.9%	1.1%
3 words	3.8%	3.9%	2.9%	2.5%
2 words	9.4%	7.5%	3.4%	4.6%
1 word	11.1%	10.4%	7.1%	7.1%
0 words	66.1%	71.2%	82.0%	80.4%
Average	**1.3**	**1.1**	**0.6**	**1.0**

On average, 1.3 words were not understood in the worst version; while in the original version, the average is only 0.6. We computed the Spearman correlation coefficient between variable S3 and the version variable, which can be treated on an ordinal scale, as described in Section 6.2.5. The correlation is relatively small (-0.14), but it is still statistically significant ($p < 0.01$). This tendency can also be intuitively observed in the share of respondents who reported finding none of the words difficult to understand: Version -2 (66.1%), Version -1 (71.2%) and Version 0 (82.0%).

Also for variable S3, we checked its interaction with three control variables and found that those who were more educated ($t = -4.7$; $p < 0.01$) and native speakers ($t = -4.2$; $p < 0.01$) reported fewer words that were not understood. On the other hand, there were no significant differences between the two genders ($t = 2.5$; $p = 0.46$). In addition, we also tried to examine the interaction in more detail with a Poisson regression model (because variable S3 is a count variable), where S3 was the dependent variable and version, education, language and gender were independent variables. The model, however, explained very little and there were no significant effects.

We can argue that subjective evaluations demonstrated that variation in wording actually has certain effects. While there were no effects on general enjoyment, the versions using synonyms with higher wording frequencies showed that respondents did detect difficulties, particularly in Version -2 (worst) but also in Version -1 (worse).

6.3 Comparing wording frequencies to results of the experiment

We can summarise this chapter with comparisons of the results of the split-ballot experiment with corpora frequencies. Table 6.21 lists all single-word changes, arranged so that the more frequent word (according to enTenTen) is on the right. The ratio (wf2/wf1) between the high- (wf2) and low-frequency wording (wf1) was computed for all cases. In the last column, we classified wording frequencies into three bands according to their enTenTen frequency: Low (up to five digits), Medium (six digits) and High (seven digits or more). About 30% of words are in the low band (rare words with frequencies ranging from 506 to 73,728), 40% in the medium band (ranging from 111,269 to 839,351), and 30% in the high band (ranging from 1,284,950 to 24,8380,312).

Table 6.21: Ratio difference between single frequencies and frequency bands

Single words (w1-w2)**	Version	wf1	wf2	wf2/wf1	Band change*
P0. careful-cautious (adj)	-2 to -1	180225	833568	4.6	M-M
P1. menace- threat (n)	-1 to 0	50138	666925	13.3	L-M
P1.II subduing-reducing (v)	-2 to -1	4778	782021	163.7	L-M
P2. apprehensive-worried (adj)	-1 to 0	18993	324153	17.1	L-M
P2. worried-concerned (adj)	0 to 1	324153	776023	2.4	M-M
P2.II attempt-attack (n)	-2 to -1	2804183	3008661	1.1	H-H
P3. vindicated-justified (adj)	-1 to 0	8863	111269	12.6	L-M
P3.II torture-torturing (n, v)	0 to 1	21926	351772	16.0	L-M
P3.III hypothetical-suspected (adj)	-2 to -1	66564	272533	4.1	L-M
P3.IV acquire-gain (v)	-2 to -1	1613498	3091890	1.9	H-H
P4.0 periodic-occasional (adj)	-1 to 0	125855	174986	1.4	M-M
P4.0.II constitute-be part of (v)	-2 to -1	433976	13662023	31.5	M-H
P4.1 probability-chances (n)	-1 to 0	222141	482356	2.2	M-M
P4.1 chances-risk (v)	0 to 1	482356	2526058	5.2	M-H
P4.1II be anxious-worry (v)	-2 to -1	271186	2080814	7.7	M-H
P4.2-3 disposed-sympathetic (adj)	-1 to 0	7501	12456	1.7	L-L
P4.2-3 sympathetic-support (adj, v)	0 to 1	12456	6192586	497.2	L-H
P4.2II broaden-extend (v)	-2 to -1	140758	1470710	10.4	M-H
P4.3II permitted-allowed (v)	-2 to -1	370033	1939253	5.2	M-H
P4.3II investigate-search (v)	-2 to -1	839351	5607506	6.7	M-H
P4.3IV tribunal-court (n)	-2 to -1	135449	4223826	31.2	M-H
P4.4 curtailing-restricting (v)	-1 to 0	4859	39914	8.2	L-L
P4.4 restricting-limiting (v)	0 to 1	39914	116859	2.9	L-M
P4.1II excessively-too far (adv)	-2 to -1	73728	12471272	169.2	L-H
P4.5 assembling-collecting (v)	-1 to 0	32420	208996	6.4	L-M
P4.5 collecting-gathering (v)	0 to 1	208996	340008	1.6	M-M
P4.1II similar-like (adj)	-2 to -1	3331139	35727539	10.7	H-H
P4.6a overpowering-overwhelming (adj)	-1 to 0	31884	358582	11.2	L-M
P4.6b overcome-defeat (v)	0 to 1	703885	720464	1.0	M-M
P4.6II manner-way (n)	-2 to -1	1399695	24642664	17.6	H-H
P5a. demanded-ransom money (n)	0 to 1	506	856	1.7	L-L
P5a. ransom money-ransom (n)	-2 to -1	856	27098	31.7	L-L
P5b. sureties-hostages (n)	-1 to 0	1570	15729	10.0	L-L
P6. inclined-prone (adj)	-1 to 0	118399	140703	1.2	M-M
P6.II. comes-is (v)	0 to 1	5138057	248380312	48.3	H-H
P6.III. perspective-views (n)	-2 to -1	1284950	1383554	1.1	H-H
P7. boost-encourage (v)	-1 to 0	603199	1311056	2.2	M-H
P7. encourage-promote (v)	0 to 1	1311056	1488081	1.1	H-H
P7.II worshippers-believers (n)	-2 to -1	22985	231200	10.1	L-M
P8. preoccupied-concerned (v)	-1 to 0	21312	776023	36.4	L-M
P9. reckon-consider (v)	-1 to 0	173879	7636685	43.9	M-H

**Band: L - Low (up to 5 digits in enTenTen); M - Medium (6 digits); H - High (7 digits)*

*** Part of speech: adj – adjevtive; adv – adverb; n – noun; v - verb*

In total, there were five changes from low to low (L-L), 11 from low to medium (L-M), two from low to high (L-H), seven from medium to medium (M-M), nine from medium to high (M-H), and seven from high to high (H-H). We can also observe a relation between the band change and frequency ratio (e.g., the highest ratio, sympathetic–support, belongs to the L-H band).

Table 6.22 shows which cases belong to which band. In addition, cases for which we found any effects on response quality indicators (RD, DK, AQ) are highlighted in bold.

Table 6.22: Frequency bands and effect of response quality indicators

Original	Changed to Low	Changed to Medium	Changed to High
Low	**Disposed-Sympathetic (RD, AQ)** Curtailing-Restricting Demanded money-Ransom money Ransom money-Ransom Sureties-Hostages	Menace – Threat Subduing – Reducing **Apprehensive-Worried (RD)** **Vindicated-Justified (RD, DK)** Torture-Torturing Hypothetical-Suspected Restricting-Limiting **Assembling-Collecting (RD)** Overpowering-Overwhelming Worshippers-Believers **Preoccupied-Concerned (RD)**	**Sympathetic-Support (RD, AQ)** Excessively-Too far
Medium	-	Cautious-Careful **Worried-Concerned (RD)** **Periodic-Occasional (RD, AQ)** Probability-Chances **Collecting-Gathering (RD)** Overcome-Defeat **Inclined-Prone (RD)**	Constitute-Part Chances-Risk Anxious-Worry Broaden-Extend Permitted-Allowed Investigate-Search Tribunal-Court Boost-Encourage Reckon-Consider
High	-	-	Attempt-Attack Acquire-Gain Similar-Like Manner-Defeat Comes-Closer Perspective-Views Encourage-Promote

Split-ballot experiment results: RD = response distribution; DK = don't know; AQ = acquiescence

Response quality effects were observed for one of the L-L changes, four of the L-M changes, one of the L-H changes, and four of the M-M changes.

Beyond single frequencies, Table 6.23: Ratio difference between the two string frequenciesTable 6.23 lists the frequencies for strings of words and their ratios (wf2/wf1). The last column includes response differences between versions.

Table 6.23: Ratio difference between the two string frequencies

String w1/w2	Version	wf1	wf2	wf2/wf1	SplitB*
P0. careful/cautious in dealing	-2 to -1	85	147	1.7	
P1. menace/ threat of terrorism	-1 to 0	126	2209	17.5	
P1.II subduing/ reducing the threat	-2 to -1	2	602	301.0	
P2. how apprehensive/ worried/	-1 to 0	30	824	27.5	RD
P2. how worried/concerned	0 to 1	824	963	1.2	
P2.II terrorist attempt/attack	-2 to -1	404	65935	163.2	
P3. ever vindicated/justified	-1 to 0	3	227	75.7	RD, DK
P3.II use of torture against/torturing suspected	0 to 1	14	42	3.0	
P3.III hypothetical/suspected terrorists	-2 to -1	5	3749	749.8	
P3.IV acquire/gain information	-2 to -1	3994	5891	1.5	
P4.0 periodic/occasional acts	-1 to 0	13	111	8.5	RD, AQ
P4.0.II constitute/be part of life	-2 to -1	102	878	8.6	
P4.1 probability/chances of attack	-1 to 0	28	49	1.8	
P4.1 chances/risk of attack	0 to 1	49	483	9.9	
P4.1II often anxious/worry	-2 to -1	378	2415	6.4	
P4.2-3 disposed to/sympathetic to terrorists	-1 to 0	0	18	-	RD, AQ
P4.2-3 sympathetic to/support terrorists	0 to 1	18	458	25.4	(4.3)
P4.2II broaden/extend to groups	-2 to -1	0	39	-	
P4.3II permitted/allowed to search	-2 to -1	69	377	5.5	
P4.3II investigate/search the houses	-2 to -1	7	161	23.0	
P4.3IV tribunal/court order	-2 to -1	480	70690	147.3	
P4.4 curtailing/restricting liberties	-1 to 0	6	7	1.2	
P4.4 restricting/limiting liberties	0 to 1	4	7	1.8	
P4.1II excessively/too far	-2 to -1	30	217221	7240.7	
P4.5 assembling/collecting information	-1 to 0	83	4417	53.2	RD
P4.5 collecting/gathering information	0 to 1	4417	7542	1.7	
P4.1II people similar to/like me	-2 to -1	66	26798	406.0	
P4.6a overpowering/overwhelming force	-1 to 0	194	2573	13.3	
P4.6b overcome/defeat terrorism	0 to 1	52	774	14.9	
P4.6II manner/way to defeat	-2 to -1	7	2806	400.9	
P5a. demanded/ransom money for hostages	0 to 1	0	1	-	
P5a. ransom money/ransom for hostages	-2 to -1	1	9	9.0	RD
P5b. money for surities/hostages	-1 to 0	0	1	-	
P6. inclined/prone to violence	-1 to 0	39	452	11.6	RD
P6.II. comes/is closer	0 to 1	4681	29237	6.2	
P6.III. your own perspective/views	-2 to -1	1042	1105	1.1	
P7. boost/encourage violence	-1 to 0	2	517	258.5	
P7. encourage/promote violence	0 to 1	517	1002	1.9	

P7.II among its worshippers/believers	-2 to -1	16	50	3.1	
P8. preoccupied/concerned about extremism	-1 to 0	0	3	-	RD
P9. reckon/consider yourself	-1 to 0	85	19113	224.9	

* Split-ballot experiment results: RD = response distribution; DK = don't know; AQ = acquiescence

String frequencies are lower than single frequencies and computed ratios differ. We may also add that in some cases with zero frequencies, the ratios could not be defined. However, in general, there is a clear, positive correlation between ratios computed on string and single frequencies (Spearman's rho = 0.42; p = 0.01). String frequencies and the computed ratios should be interpreted with caution, as they also depend on the length of the selected context. Therefore, we did not form frequency bands for string frequencies.

Let us compare string ratios with results of the split-ballot experiment for individual items. For eight items (P2, P3, P4.0, P4.3, P4.5, P5, P6 and P8), there were differences in response distributions; and for one of them, there was also a difference in the percentage of DK responses that was observed (P3). In addition, for two items, there was more acquiescence (P4.0 and P4.3). However, the word frequency change ratio is not particularly high for any of them. Considering Table 6.22, all of the effects occurred in cases where the original word was in the low or medium frequency band, so it seems that the frequency band and the related absolute frequency of the original wording is a more important factor than the relative word frequency change ratio.

Next, based on versions where the changes took place, we computed the median change from Version 0 to each of the other three changed versions, both for string frequencies and for single frequencies (Table 6.24). We also added the comparisons of the maximum values of the ratios.

Table 6.24: Number of changes, median and maximum for ratios of wording frequencies

Pages and items	Version -2	Version -1	Version 0	Version 1
Total number of changes (from V0)	34	16	-	11
Median ratio (from V0) for string frequencies	13.3	15.4	-	3.0
Median ratio (from V0) for single word	8.2	8.2	-	2.6

Pages and items	Version -2	Version -1	Version 0	Version 1
frequencies				
Maximum ratio (from V0) for string frequencies	7,240	258	-	25.4
Maximum ratio (from V0) for single word frequencies	170	44	-	497

We can observe that the main difference between Version -1 and Version -2 is the number of changes (34 vs. 16) as well as the extremes in the ratios (7,240 vs. 258), while the median level in the changes of the ratios was not that different.

Version 1, on the other hand, has relatively small overall effects in the variation of the ratios. The median values (3.0 and 2.6) suggest that the frequencies used in Version 1 were around three times higher than those used in the original Version 0.

Of course, we should not forget that there were differences between the four versions in other aspects as well (response time, drop-outs, satisfaction, response patterns, etc.).

6.4 Discussion

In this chapter, we analysed differences among the four versions of the questionnaire (which differ with respect to corpora frequencies of the corresponding wordings) for five indicators of response quality. Let us summarise the essential findings:

- The drop-out rate in the worst version (-2) was 17%, while in the other three versions it was about 13%. Drop-outs mostly occurred at the fifth question, which has a matrix format; also at that point, the worst version has about a 4-percentage point increase in drop-outs than the other three versions.

- The more difficult the version, the longer it took respondents to answer; but after removing outliers, the response times are significantly higher only for the worst version (-2).

- There were almost no effects found with respect to the level of DKs, except for one item where the two versions that used the unfamiliar wording ‘vindicated’

had significantly more DK answers (11%) than the two versions which used the wording ‘justified’ (8%).

- For acquiescence, measured as the agreement with all seven sentences in the matrix question, there was a significant effect for two of the items, but with somewhat opposing results. In one case, there was more agreement when using the less frequent wordings than the more frequent wordings, which goes contrary to our expectations; while in the others, there was more agreement when replacing the original wording with the less frequent word – which was the expected direction – but also when the wording was improved. Thus, we should refrain from making some clear conclusions on the effect of wording changes on acquiescence.

- The analysis of subjective evaluations of questionnaire difficulty shows that respondents to the worse and particularly the worst versions actually noticed the problems which arose from the lower frequency of words used. In fact, they found more question meanings difficult to interpret than those who responded to the original and improved versions, and they also reported higher numbers of words that were at least a little difficult to understand.

- It seems that men, non-native speakers, and those with a lower education found the questionnaire more difficult and reported a higher number of words that were at least a little difficult to understand. The relationship between these indicators and questionnaire version was systematically controlled for these three background characteristics. On one hand, there was no interaction with gender, education, and language for the association between version and the number of difficult words. On the other hand, for questionnaire difficulty, there was some interaction: When controlling for gender, the worst version was still found to be more difficult, but the association was weaker for both genders. Both native speakers and those with a higher education were affected by changes in word frequency; while for non-native speakers and those with a lower education, there was no version effect – probably because they found the questionnaire to be difficult regardless of the version.

- When studying wording frequencies in more detail, it appears that for those cases where there are significant differences in response distributions and/or response quality between different versions, they are not primarily due to the relative change ratio of the improvement but more a factor of the frequency band and absolute frequency of the original single word.

Since the alternative wordings are all synonymous, we did not expect many differences in response distributions; however, particularly in the ninth question, there were significant differences between the versions that used the wording ‘concerned’ and the versions that use ‘preoccupied’. This will be further discussed in the final interpretation, where the results from expert evaluations and cognitive interviews are also included.

Based on our experiment, the basic conclusion is that low-frequency wordings have a significant cumulative effect on response quality, at least when there are enough wording changes within survey questions throughout the questionnaire. If we make only minor changes, such as in Version -1 and Version 1, only a small effect can be observed. In fact, most of the significant differences are attributable to the worst Version (-2), which has more than twice the number of worsened wordings than the worse Version (-1).

7 Conclusions

Survey data collection is the prevailing method in quantitative social science research, and writing good survey questions is an important part of ensuring the high quality of collected data. Within this context, deciding on the optimal question wording among several alternatives is one of the most complex issues in the process of questionnaire development. Several methods of evaluating survey questions exist to improve these aspects, but they are often very complex and resource-consuming. As in many areas, attempts have been made to make the process of question evaluation and its improvement simpler and more accessible. One of the directions that can be taken to achieve this goal is based on linguistic resources, which have been underutilised in survey methods research.

In this dissertation, we studied how linguistic resources can be used to improve the process of pre-testing survey questionnaires. After the introduction and outline of the study (Chapter 1), we presented in Chapter 2 the basic concepts of corpus linguistics and semantics and discussed how text corpora and other linguistic resources can be used to detect unfamiliar wordings and find alternatives; we also discussed previous applications of text corpora in survey methods research. In addition, we introduced the traditional approaches of question evaluation, both qualitative (cognitive interviewing, expert evaluation) and quantitative (split-ballots).

In Chapter 3, we presented a preliminary pilot study: a wording experiment done on the English and Slovenian versions of a questionnaire used for evaluating international student exchange programmes, where two versions of the same questionnaire were developed, one with low-frequency and one with high-frequency wordings. This study helped operationalise the main empirical research.

In Chapter 4, we designed and conducted expert reviews to evaluate two case studies: a set of questions from the WageIndicator questionnaire on wages and working conditions, and a selection of PEW research questions on the topic of terrorism. Language resources were used to select the cases with alternative wordings, and text corpora frequencies were calculated.

In Chapter 5, we presented the results of online cognitive interviews. We compared the performance of the two versions of the PEW questions, where alternative wordings were used (i.e., synonyms with different corpora frequencies). Correspondingly, we observed the potential differences in the meaning and understanding of alternative wordings.

In Chapter 6, we presented the results of our main empirical study: a split-ballot experiment on essentially the same selection of PEW questions already used in expert reviews and cognitive interviews (with a few additional items). Four versions of the same questionnaire were compared: the original PEW questions, an improved version of the same questions, and two worsened versions that differed in the number of changes. Five data quality indicators were observed: drop-outs, response times, 'don't know' responses, acquiescence and subjective evaluations of respondents.

7.1 Main findings

Let us briefly summarise the main findings of this dissertation.

- **Preliminary pilot experiment**: Although there were some methodological limitations, the first study confirmed the basic findings of Lenzner (2011) that word frequencies can affect some aspects of question comprehension and response quality. However, due to the differences in methodological approaches, some effects found in our study were different. While Lenzner found an impact on response times, we found effects on drop-out rates and on the subjective perception of response burden: a novel finding in our study and one which contributes to a broader area of social science methodology. Nevertheless, both studies found no effect on item nonresponse and satisficing. The preliminary study also identified some key factors for potential inclusion in experimental designs in future studies; it also suggested that strings of words which are available within some of the existing resources should be used in future study designs. This was accomplished in the other three studies that we conducted for this dissertation.
- **Comparison of corpora frequencies and expert reviews**: The main finding of the second study is that the approach based on text corpora – which is semi-automated and inexpensive – can to a considerable extent replace lengthy and resource-demanding expert evaluations. In fact, in more than half of the evaluated cases,

corpora frequencies matched expert evaluations of the appropriateness of a question wording in a certain context. In some other cases, however, there were certain divergences between the two approaches: In most cases, this was due to the string frequencies being too low to make effective comparisons; in a few cases, however, we could not provide an explanation. Moreover, in many cases, there were substantial variations between experts, particularly between native and non-native speakers. These results are in line with what Graesser (2006) and Olson (2010) found when comparing expert reviews and the results of QUAID software.

- **Cognitive interview evaluation:** Two versions of the same questions were compared in an online split-ballot cognitive interview that used the paraphrasing and definition techniques to study how respondents understand wording alternatives with the same meaning but with different wording frequencies. In most cases, when presented with a low-frequency wording, respondents used its high-frequency alternative to define it or to paraphrase the whole item. Another finding is that there was much more variation in participants' answers in certain cases of high-frequency wordings, which indicates that these wording alternatives might have less clear and more ambiguous meanings. For most cases, this is also reflected in the number of WordNet senses, which is higher for more frequent words. In fact, according to the Krylov Law of Polysemy, more frequent words are more likely to have more meanings.
- **Main experiment:** This study observed the data quality indicators across the four versions of the questionnaire (original, improved, worse and worst) and basically confirmed that the familiarity of question wordings measured with corpora frequencies can affect the response quality of the survey data. Similar to Lenzner et al. (2010), we also confirmed a longer response time for low-frequency wordings. Moreover, we confirmed the difference in drop-out rates, which Lenzner et al. (2010) were not able to achieve. For one of the items, we found a lower number of non-substantive (DK) responses, as was also found by Lenzner (2012). However, the effects were mostly small and can be observed only in cases where there are several wording changes made (each with significant differences in corpora frequencies), such as in the case of our worst version. On the other hand, making only a small number of wording changes would not produce enough differences in most of the survey response data quality indicators. The majority of the observed effects relate to the two worsening conditions, while there is less difference between

the original and improved version. The indicators for which the effects were the strongest are drop-outs and two of the subjective indicators of response quality, questionnaire difficulty, and the reported number of words not understood. For the last two, we also controlled for the interaction with three background variables: gender, education and language. No interaction was found for the number of words, while all three interacted in the association between questionnaire version and perception of difficulty: gender weakened the effect, while for education and language we observed that it affected only native speakers and those highly educated; non-native speakers and those with a lower education, on the other hand, found the questionnaire difficult regardless of which version they were responding to.

Table 7.1 provides a synthesis of the findings for the selected PEW items, as it shows the results from text corpora (i.e., corresponding frequency ratios; see Section 6.3), expert reviews (see Section 4.4.2), cognitive interviews (H-high, M-median or L-low match of the wordings, which denote the level of synonymy; see Section 5.2.2), and the significant effects from the survey experiment (see Section 6.3).

Table 7.1: Comparison of quantitative and qualitative results for all available items

String (W1/W2)	WF ratio	Expert reviews	Cognitive interviews	Split-ballot experiment
P0. careful/cautious in dealing	1.7	x	H	-
P1. menace/ threat of terrorism	17.5	menace < threat	H	-
P2. how apprehensive/ worried/	27.5	apprehensive < worried, concerned	L	RD
P2. how worried/concerned	1.2		-	
P3. ever vindicated/justified	75.7	vindicated < justified	L	RD, DK
P4.1 probability/chances of attack	1.8	probability, chances < risk	M	-
P4.1 chances/risk of attack	9.9		-	
P4.2-3 disposed to/sympathetic to terrorists	-	disposed to < sympathetic to < support	M	RD, AQ (4.3)
P4.2-3 sympathetic to/support terrorists	25.4		-	
P4.4 curtailing/restricting liberties	1.2	curtailing < restricting < limiting	L	-
P4.4 restricting/limiting liberties	1.8		-	
P4.5 assembling/collecting information	53.2	assembling < collecting, gathering	M	RD
P4.5 collecting/gathering information	1.7		-	

String (W1/W2)	WF ratio	Expert reviews	Cognitive interviews	Split-ballot experiment
P5a. demanded/ransom money for hostages	-	demanded money < ransom money	H	RD
P5a. ransom money/ransom for hostages	9.0			
P5b. money for sureties/hostages	-			
P6. inclined/prone to violence	11.6	prone, inclined	M	RD
P7. boost/encourage violence	258.5	boost < promote, encourage	L	-
P7. encourage/promote violence	1.9			
P8. preoccupied/concerned about extremism	-	preoccupied < concerned	L	RD

We can observe from the above table that there exists a certain link between the levels of the match of the alternative wordings, as high matches (i.e., good synonyms) rarely produce effects on response behaviour (i.e., significant effects in the experimental study). These effects also seem to appear more often when the ratio of the wording differences is higher.

7.2 Study limitations and potential for future research

Let us summarise the key limitations of the empirical studies in this dissertation:

- Although the items and cases for empirical cases were very carefully selected so that the effects of different wording frequencies were isolated as much as possible, the specifics of the selected examples could impact the nature of the conclusions. Related to this, a particular concern is the question of how realistic the effects are – or are they a product of exaggeration? In fact, in many cases, we were not improving questions but actually worsening them. Is the worst version realistic? Would somebody really word questions like that? Perhaps the selected case studies were already good enough, and if we would have worked with worse questions, there would have been more room for improvement. Such questionnaires certainly do exist, so the research we do is important.
- Questionnaire improvements should be tailored to the population studied. For instance, the pilot study aimed at students might have benefited from using a more specific corpus based on texts that were familiar to this population. And even when studying the general population, as we did in the main study, there are differences

between different groups of respondents. Optimally, question improvements would be tailored to their education level and other characteristics. However, in practice this is not very feasible. So, our recommendation for general populations is to try to make improvements that would benefit those who are the weakest link, while tailoring should be used in cases of specific populations (e.g., a questionnaire aimed at lawyers).

- There were some problems with paradata from the survey experiment. As time stamps per page (or per question) were not available, we could not do case-level analyses of response latencies.
- Another limitation is that in one-quarter of the cases in this study, the corresponding word frequencies were too low.

Future directions for research are related to the potential model, which would be able to predict (for synonyms) the impact of the variation in corpora frequencies on the response process and also on the survey data quality indicators. This should be done differentially and conditionally on the level of the similarity between the alternative wordings (e.g., total synonyms vs. partial synonyms). For this purpose, in addition to potentially new empirical studies, meta-analysis studies could be performed in the following directions:

- Applying the approach to various studies where other testing methods were already carried out. There are several pre-testing reports available, both for expert reviews (Graesser 2000; Olson 2010) and cognitive interviews. This makes comparisons across methods possible. In the comparison, the specific focus should be on the cross-validation of the effectiveness **of pre-testing methods for detecting wording problems** and on finding corresponding alternatives.
- Applying the approach to selected questions in the Q-bank database (English), for which cognitive interviews were already performed. The Q-bank is a collection of reports completed by different testing agencies in the US on different question topics. The approach could involve coding the results of cognitive interviews for selected studies and performing the related corpora frequency analysis.
- Taking better advantage of functionalities provided in corpus linguistic tools, such as Sketch Engine. First, this would consist of taking advantage of corpus metadata and tailoring the results to specific genres, sources, times of publication, etc. One

particular metric that would be useful to compute is the dispersion of a word across different sources. Second, in future research, we should go beyond simple concordances to also compute collocations of words in a specific context. To some extent, this is already possible using the sketch difference functionality; however, the format of the query does not currently allow choosing the specific context. This is a feature that Sketch Engine is going to add soon.

- Beyond corpus linguistics, other language diagnostics could also be considered for integration with the approach discussed in this dissertation, particularly various readability and comprehensibility indicators we mentioned in the introduction.

Systematic meta-analytic studies should also be conducted to discover key factors, mediators and effects in this complex matter (e.g., the role of the target population, the role of various languages, and the role of words with foreign origins).

We may add that, as already mentioned in the introduction (Section 1.4), we are conducting another split-ballot experiment which is currently in the data collection phase. Specifically, we are evaluating the WageIndicator questionnaire, both the English and Slovenian version. For both questionnaires, an alternative version with improved wording was developed and is being tested in the field at the time this dissertation is being submitted. This additional research will provide insight into yet another case which was already in part included in this dissertation.

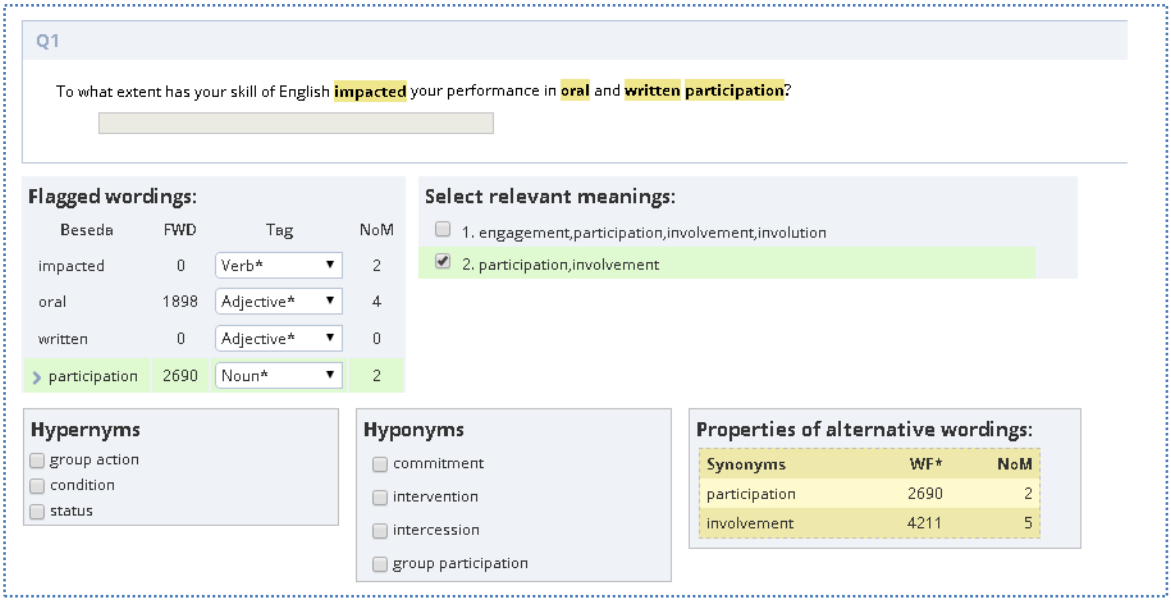
7.3 Potential for the integration of language resources into questionnaire development tools

Finally, we would like to mention a specific practical aspect of our research. Namely, one of the motivations for this study was the idea of developing a prototype tool (software) that would be able to, among other features, flag unfamiliar wordings and suggest improvements. We actually developed a pilot application which integrated the **linguistic corpora and dictionaries** for detecting word unfamiliarity and ambiguity **as well as lexical databases** as a source for looking up synonyms. We considered the options for both English and Slovenian and made a sensitivity analysis on a pilot questionnaire to check differences among them. In this step, we took into account the

incompleteness of the Slovenian lexical database, sloWNet, and performed the required manual editing.

This **prototype application** has already been integrated into the 1KA web survey software (www.1ka.si, Navigation Testing → Language Technology) so that developers of survey questionnaires can use this functionality to test the frequencies of different wordings. Figure 7.1 shows an example for one of the questions from the questionnaire for international exchange students (i.e., the study presented in Chapter 3).

Figure 7.1: Screenshot of language technology module in prototype application



In the first step, the application examines all single words in the question and flags those that exceed at least one of the three predefined parameters (word frequency threshold, number of meanings for nouns, and number of meanings for verbs). For instance, in the above example, the words ‘impacted’, ‘oral’, ‘written’, and ‘participation’ are flagged. The flagged words are also listed in a table under the question, which also lists their wording frequency (FWD), word type (Tag) and number of meanings (NoM). Note that the word type (noun, verb, adjective or adverb) is automatically detected, but users are able to change the parameters (dropdown menu) in case of an application error.

In the next step, the user can select (by clicking) any of the listed flagged words and a list of WordNet meanings will be displayed. For instance, in the above example, we clicked the word ‘participation’, for which the application found relevant meanings. However, at present, it only includes synsets without the full definitions of the words. The meanings are shown in a checkbox format, and the user has to select the intended meaning (or even several meanings) in case they are not sure or want to examine a wider array of alternatives. In addition, the user can also select from the hyponyms and hypernyms that are listed in WordNet for each individual word. We also plan to add a field that would enable the user to add additional alternatives; for example, those found in other resources such as thesauri.

In the last step, after the desired alternatives are checked, the synonyms for the selected meanings – and also hyponyms, hypernyms and other alternative words (if selected) that could be potential improvements – are displayed in a table, including two of their properties, wording frequency (WF), and number of meanings (NoM). In the above example, for instance, we selected the second meaning: ‘participation, involvement’. Finally, it is up to the user to decide, based on the results, what is the optimal wording for that context. In cases where unfamiliarity is in conflict, it is recommended that users give priority to the high word frequency criteria over the low number of meanings.

Currently, the application only works for single words. However, as revealed when examining the case studies used this dissertation, we learned that it is essential to check the context of the words (i.e., strings of neighbouring words) and not just the word frequency parameter. Thus, it is necessary to further develop the application so that it can also examine strings of words and provide collocations and word sketches. In addition, for the full exploitation of language resources and their integration into questionnaire design and the testing process, the application should also provide standard coefficients of question complexity, readability and comprehensibility, as well as spelling and grammar checkers.

Unfortunately, at the advanced level, the corresponding application would become much more complex and would require substantial resources; so, at this point, it has not yet been developed. Nevertheless, it is true that only with these extended functionalities can the application become fully useful. Still, the initial level can already help in many situations (i.e., it can help discover synonymous words with low word frequencies).

In addition, it would be worth exploring the possibilities for detecting other problematic language characteristics, such as word abstraction, double-barrelled questions, double negatives and leading questions. However, for these indicators, there seems to be no simple, automated method for suggesting alternatives, at least not at this point.

7.4 Originality of contribution

This thesis presents an important contribution to survey methodology and to a broader field of social science methodology and statistics. Within this context, a new approach for question evaluation was developed and tested. We used language resources in a different way than previously used in survey research. First, we relied on the retrieval of alternative wordings from the WordNet lexical database to form alternative wordings; second, we used text corpora to calculate the corresponding frequencies of alternative wordings (i.e., strings of words and not just single words). After that, we extensively evaluated this approach with cognitive interviews and expert judgements, as well as with survey data quality indicators from an empirical study.

Based on this work, we can confirm the main thesis of the dissertation: the above described implementation of language resources can be effectively integrated for improving the comprehension of survey questions.

From a theoretical standpoint, this thesis contributes to studies related to cognitive aspects of question development. Similarly, it also presents an important contribution to the expanded conceptualisation and understanding of the role of information-communication technologies in the survey process.

From a practical perspective, this thesis has laid the groundwork for the development of an approach that could be used as a supplement to other questionnaire evaluation methods in the development and pre-testing stage of survey questionnaire design. For the English language, there were already some previous developments that we have upgraded; while for the Slovenian language, this is the first attempt in this direction. This contribution is particularly important in contemporary times, when powerful and inexpensive tools are available to enable users to create surveys easily and en masse. Thus, technology increasingly enables the production of low quality

questionnaires. The work in this dissertation is an attempt to offer tools that could help improve the quality of at least one aspect of survey questionnaires.

The proposed method thus presents an improvement to the field of questionnaire design, but may also have applications outside the narrow field of wording survey questions. In fact, the problem of selecting the most comprehensible wording extends far beyond the field of survey methodology.

References

Akkerboom, Hans and Francine Dehue. 1997. The Dutch Model of Data Collection Developments for Official Surveys. *International Journal of Public Opinion Research* 9 (2): 126–145.

Blasius, Jörg and Jurgen Friedrichs. 2009. The Effect of Phrasing Scale Items in Low-Brow or High-Brow Language on Responses. *International Journal of Public Opinion Research* 21(2): 235–247. doi:10.1093/ijpor/edp018

Bradburn, Norman M. 1978. *Respondent burden*. In Proceedings of the Survey Research Methods Section of the American Statistical Association: 35–40.

---, Norman M. and Seymour Sudman. 1983. *Asking questions: A practical guide to questionnaire design*. San Francisco, CA: Joessey–Bass.

Broadbent, Donald E. 1967. Word-frequency effect and response bias. *Psychological review,* 74 (1): 1–15. doi:10.1037/h0024206

Burnard, Lou. 1995. *Users Reference Guide British National Corpus Version 1.0.* Oxford: Oxford University Computing Services.

Cantril, Hadley. 1944. *Gauging Public Opinion*. Princeton, NJ: Princeton University press.

Cerar, Teja, Nina Konavec, and Valentina Hlebec. 2011. Uporaba ekspertnih shem za kvalitativno testiranje anketnih vprašalnikov. *Teorija in praksa* 48 (2): 393–410.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coltheart, Max. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology. Section A: Human Experimental Psychology* 33 (4): 497–505.

Converse, Jean M. and Stanley Presser. 1986. *Survey questions: Handcrafting the standardized questionnaire*. Thousand Oaks, CA: Sage.

Couper, Mick P., Reginald P. Baker, Jelke Bethlehem, Cynthia Z. F. Clark, Jean Martin, William L. Nicholls, and James M. O'Reilly, eds. 1998. *Computer Assisted Survey Information Collection*. Hoboken, NJ: Wiley Series in Probability and Statistics.

---, Mick P. and Frauke Kreuter. 2013. Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176: 271–286.

Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25 (4): 447–464. doi: 10.1093/llc/fqq018

Duffy, Susan A., Robin K. Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of memory and language* 27(4): 429–446.

Duncan, Otis Dudley and Howard Schuman. 1980. Effects of Question Wording and Context: An Experiment with Religious Indicators. *Journal of the American Statistical Association* 370 (75): 269–275.

Fazio, Russell H. 1990. A practical guide to the use of response latency in social psychological research. *Review of Personality and Social Psychology* 11: 74–97.

Fellbaum, Christiane D., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fišer, Darja. 2009. *Izdelava slovenskega semantičnega leksikona z uporabo eno– in večjezičnih jezikovnih virov*. Doctoral Dissertation. Ljubljana, Slovenia: Faculty of Arts.

Flesch, Rudolf. 1943. Marks of readable style: a study in adult education. *Teachers College Contributions to Education* 897: 9–69.

Graesser, Arthur C., Tina Kennedy., Peter M. Wiemer–Hastings, and Victor Ottati. 1999. The use of computational cognitive models to improve questions on surveys and questionnaires. In Sirken, Monroe G., Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur and Roger Tourangeau, Eds., *Cognition and Survey Research:* 199–216. Hoboken, NJ: Wiley Series in Probability and Statistics.

---, Arthur., Katja Wiemer–Hastings, Paul Wiemer–Hastings and Roger Kreuz. 2000. *The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices.* In Proceedings of the Section on Survey Research Methods of the American Statistical Association: 459–64.

---, Arthur, Zhiqiang Cai, Max M. Louwerse and Daniel Frances. 2006. Question Understanding Aid (QUAID) A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly* 70 (1): 3–22.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey methodology, 2nd edition.*. Hoboken, NJ: Wiley.

Hedlin, Dan, Trine Dale, Gustav Haraldsen, and Jacqui Jones. 2005. *Developing Methods for Assessing Perceived Response Burden*. Eurostat. Retrieved from: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DEVELOPING%20METHODS%20FOR%20ASSESSING%20PERCEIVED%20RESPONSE%20BURD.pdf (January 23 2014).

Heerwegh, Dirk. 2003. Explaining response latencies and changing answers using client side paradata from a web survey. *Social Science Computer Review* 21 (3): 360–373.

Holbrook, Allyson L., Jon A. Krosnick, D. Moore, and R. Tourangeau. 2007. Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly* 71:325–48.

Howes, Davis, H. and Richard L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology* 41 (6): 401–410.

Inhoff, Albrecht W. and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics* 40 (6): 431–439.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. *The TenTen corpus family*. 7th International Corpus Linguistics Conference. Available at: https://www.sketchengine.co.uk/wp-content/uploads/The_TenTen_Corpus_2013.pdf (February 1 2016).

Johns, Tim. 1991. Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal* 4: 1–16.

Jurafksy, Daniel (2003). Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, Eds. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Kalton, Graham, Martin Collins and Lindsay Brook. 1978. Experiments in Wording Opinion Questions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 2 (27): 149–161.

Kilgariff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell. 2004. *The Sketch Engine*. 11th Euralex International Congress. Lorient, FR.

Kincaid, Peter J., Robert P. Fishbourne, Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, For Coung and Flesch Reading Ease Formula) for navy enlisted personnel*. Memphis, TN: Chief of Naval Technical Training.

Krosnick, Jon A. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5 (3): 213–236.

---, Jon A, Sowmya Narayan, Wendy R. Smith. 1996. Satisficing in surveys: Initial evidence. In Braverman, Marc. T and Jana Kay Slater, Eds., *Advances in Survey Research*: 29–44. San Francisco: Jossey–Bass.

---, Jon A. and Leandre R. Fabrigar. *The handbook of questionnaire design*. New York, NY: Oxford University Press (forthcoming).

Krylov, Jury K. 1982. *Ob odnoj paradigme lingvističeskich raspredelenij.* In: Trudy po lingvostatistike 8: Lingvostatistika i vyčisliteč'naja lingvistika, 80–102.

Kučera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London, UK: Longman.

Lenzner, Timo, Lars Kaczmirek, and Alwine Lenzner. 2010. Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology* 24 (7): 1003–1020. doi:10.1002/acp.1602

---, Timo. 2011. *A Psycholinguistic Look at Survey Question Design and Response Quality*, Doctoral dissertation. Mannheim, DE: Universität Mannheim.

---, Timo, Lars Kaczmirek, and Mirta Galesic. 2011. Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research* 23 (3): 361–373. doi: 10.1093/ijpor/edq053

---, Timo. 2012. Effects of survey question comprehensibility on response quality. *Field Methods*, 24 (4): 409–428. doi: 10.1177/1525822X12448166

Lessler, Judith and Barbara Forsyth. 1996. A Coding System for Appraising Questionnaires. In Schwarz, Norbert and Seymour Sudman, Eds., *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, 259–292. San Francisco: Josey–Bass.

Lin, Lu-Fang. 2011. Gender Differences in L2 Comprehension and Vocabulary Learning and the Video-based CALL Program. *Journal of Language Teaching and Research* 2 (2): 295–301.

Logar Berginc, Nataša and Simon Krek. 2012. New Slovene Corpora within the Communication in Slovene Project. *Philological Studies* LXIII: 197–208.

Lozar Manfreda, Katja, Zenel Batagelj and Vasja Vehovar. 2002. Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer Mediated Communication* 7 (3): 0.

Madans, Jennifer, Kristen Miller, Aaron Maitland and Gordon Willis. 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality*. Hoboken, NY. Wiley.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. 1993. *Introduction to WordNet: An On–line Lexical Database*. Retrieved from: http://wordnetcode.princeton.edu/5papers.pdf (November 15 2013).

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 11: 39–41.

Mohorko, Anja and Hlebec, Valentina. 2013. Razvoj kognitivnih intervjujev kot metode pretestiranja anketnih vprašalnikov. *Teorija in praksa* 50 (1): 62–95.

---, Anja. 2015. *Ovrednotenje tehnik kognitivnega intervjuja kot metode za predtestiranje anketnih vprašalnikov*. Doctoral dissertation. Ljubljana, SI: Faculty of Social Sciences.

Nation, Paul and Robert Waring. 1997. Vocabulary size, text coverage and word litsts. In Norbert Schmitt and Michael McCarthy, Eds. *Vocabulary: Description, Acquisition and Pedagogy:* 6–19. Cambridge, UK: Cambridge University Press.

Olson, Kristen. 2010. An examination of questionnaire evaluation by expert reviewers. *Field Methods* 22 (4): 295–318.

---, Kristen and Jolene D. Smyth. 2015. The Effect of CATI Questions, Respondents, and Interviewers on Response Time. *Journal of Survey Statistics and Methodology* 3 (3): 361–396.

Peytchev, Andy. 2009. *Survey Breakoff*. Public Opinion Quarterly 73 (1): 74–97.

Presser, Stanley, Jennifer M. Rothgeb, Mick P. Couper, Judith Lessler, Elizabeth Martin, Jean Martin and Eleanor Singer. 2004. *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: Wiley.

Rasinski, Kenneth. 1989. The Effect of Question Wording on Public Support for Government Spending. *Public Opinion Quarterly* 53 (3): 388–394.

Rug, Donald. Experiments in Wording Questions II. *Public Opinion Quarterly* 5 (1): 91–92.

Saris, Willem and Irmtraud Gallhofer. 2007. *Design, evaluation, and analysis of questionnaires for survey research.* Hoboken, NY: Wiley Series in Survey Methodology.

---, Willem. 2012. *Discussion Evaluation Procedures for Survey Questions.* Journal of official statistics, 28 (4): 537–551.

Schaeffer, Nora Cate and Dykema, Jennifer. 2011. Questions for Surveys Current Trends and Future Directions. *Public opinion quarterly* 75 (5): 909–961.

Schuman, Howard P and Stanley S. Presser. 1981. *Questions and answers in attitude surveys*. San Diego, CA: Sage publications.

Schwarz, Norbert. 2007. Cognitive aspects of survey methodology. *Applied Cognitive Psychology* 21 (2): 277–287.

Sheatsley, Paul B. 1983. Questionnaire construction and item writing. In Rossi, Peter H., James D. Wright and Andy B. Anderson, Eds., *Handbook of Survey Research*. Orlando, FL: Academic Press.

Sirken, Monroe G., Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur and Roger Tourangeau, Eds. 1999. *Cognition and Survey Research*. Hoboken, NJ: Wiley Series in Probability and Statistics.

Smith, Tom W. 1987. That Which We Call Welfare By Any Other Name Would Smell Sweeter: An Analysis of Question Wording on Response Patterns. *Public Opinion Quarterly* 51: 75–83.

Snijkers, Ger J. M. E. 2002. *Cognitive laboratory experiences: on pre–testing computerised questionnaires and data quality*: Ervaringen met vragenlabonderzoek: over het pre–testen van gecomputeriseerde vragenlijsten en datakwaliteit. Doctoral dissertation. Utrecht, NL: Utrecht University.

Thomas, James. 2016. *Discovering English with Sketch Engine*, 2nd edition. Versatile.

Tijdens, Kea G. and Brian Fabo. 2014. *Codebook WageIndicator web survey on work and waged*. Amsterdam: WageIndicator Data Report. Available at: http://www.uva-aias.net/51 (March 1 2015).

Tijdens, Kea G. and Paulien Osse. 2015. *WageIndicator continuous web-survey on work and ages*. Amsterdam: University of Amsterdam/AIAS and WageIndicator Foundation. Available at: http://www.uva-aias.net/90 (March 1 2015).

Tourangeau, Roger. 1984. *Cognitive science and survey methods* (Vol. 73). Washington, DC: National Academy Press.

Tourangeau, Roger, Lance J. Rips, Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge, MA: University Press.

Vehovar, Vasja, Petrovčič, Andraž and Slavec, Ana. 2014. e–Social Science Perspective on Survey Process: Towards an Integrated Web Questionnaire Development Platform. In Engel, Uwe, Ben Jann, Peter Lynn, Annette Scherpenzeel and Patrick Stugris, Eds. *Improving Survey Methods*, Chapter 15. New York, NY: Routledge.

Wei, Xuemei. 2014. An Empirical Study of Gender Effect on L2 Vocabulary Acquisition Strategies. *Studies in Literature and Language* 9 (3): 86–93.

Willis, Gordon B., Susan Schechter and Karen Whitaker. 1999. *A comparison of cognitive interviewing, expert review, and behavior coding: What do they tell us?* In Proceedings of the Section on Survey Research Methods, American Statistical Association: 28–37.

---, Gordon B. 2005. *Cognitive Interviewing. A Tool for Improving Questionnaire Design.* Thousand Oaks, CA: Sage Publications.

Yan, Ting, Frauke Kreuter, and Roger Tourangeau. 2012. Evaluating Survey Questions: A Comparison of Methods. *Journal of Official Statistics* 28 (4): 503–529.

Appendix A. Pilot survey questionnaire

*Slovenian version can be previewed at:
https://www.1ka.si/izmenjavaLJ&preview=on&mobile=0&disableif=1&pages=all

Dear student,

thanks for agreeing to participate in our survey. The questionnaire consists of 32 questions and will only take you about 12 minutes to fill in. Please click "Next page" at the bottom of the page to start.

Q34 - In which country is your home university based? Please select. [dropdown list]

Q35 - Did you enroll the exchange programme at the University of Ljubljana as a constituent of your undergraduate or postgraduate studies?

◌ Undergraduate
◌ Postgraduate (masters, doctoral)

Q36 - In which subject field would you categorise your exchange program?

◌ Natural sciences (mathematics, computer and information sciences, physical sciences, chemical sciences, earth and related environmental sciences, biological sciences, other national sciences)
◌ Engineering and technology (civil engineering, electrical engineering, electronic engineering, information engineering, mechanical engineering, chemical engineering, materials engineering, environmental engineering, environmental biotechnology, industrial biotechnology, nano-technology, other engineering and technologies)
◌ Medical and health sciences (basic medicine, clinical medicine, health sciences, health biotechnology, Other medical sciences)
◌ Agricultural sciences (agriculture, forestry, and fisheries, animal and dairy science, veterinary science, agricultural biotechnology, other agricultural sciences)
◌ Social sciences (psychology, economics and business, educational sciences, sociology, law, political science, social and economic geography, media and communications, other social sciences)
◌ Humanities and arts (history and archaeology, languages and literature, philosophy, ethics and religion, arts, history of arts, performing arts, music, other humanities)
◌ Interdisciplinary field - please specify:

Q37 - Given the study prerequisites at the University of Ljubljana, did you begin the exchange program with adequate background knowledgefrom your home university?

◌ Inadequate knowledge
◌ Just adequate knowledge
◌ More than adequate knowledge

Q38 - Given the study prerequisites at the University of Ljubljana, have you acquired adequate skills at your home university in terms of the following categories?

	Inadequate	Just adequate	More than adequate	Does not apply to my program
Analytical skills	◌	◌	◌	◌
Scientific writing	◌	◌	◌	◌
Oral communication skills	◌	◌	◌	◌
Critical assessment	◌	◌	◌	◌
Experimental skills within your field (fieldwork, laboratory work)	◌	◌	◌	◌
Computer skills (word processor, spreadsheets, presentations, e-mail, etc.)	◌	◌	◌	◌
Computer programming languages	◌	◌	◌	◌
Foreign language skills	◌	◌	◌	◌
Problem-solving	◌	◌	◌	◌
Teamwork	◌	◌	◌	◌
Leadership skills	◌	◌	◌	◌

Q39 - In general, how would you describe the quality of courses you took at the University of Ljubljana?

◌ Excellent
◌ Good
◌ Fair
◌ Poor
◌ Very Poor

Q40 - How much knowledge have you acquired from the courses at the University of Ljubljana?

◌ A great deal
◌ A lot
◌ A moderate amount
◌ A little
◌ Nothing

Q41 - Has your average exam mark at the exchange program been higher, approximately the same or lower in comparison to the one acquired at your home university (in last completed year)?

◌ Lower
◌ Approximately the same
◌ Higher
◌ Not Applicable - Have not received any marks yet

Q42 - Please evaluatedifferences in study circumstances between your home academic environment and your experience at the University of Ljubljana according to the next criteria.

	Much lower than at my home university	Lower	Approximately the same	Higher	Much higher than at my home university	Does not apply to my program
Amount of obligatory attendance	◌	◌	◌	◌	◌	◌
Amount of assignments per week	◌	◌	◌	◌	◌	◌
Prerequisites for registering for an exam	◌	◌	◌	◌	◌	◌

Q43 - Furthermore, evaluatedifferences in study conditions between your home academic environment and your experience at the University of Ljubljana according to the following aspects.

	Much worse than at my home university	Worse	Approximately the same	Better	Much better than at my home university	Does not apply to my program
Quality of lectures	◌	◌	◌	◌	◌	◌
Quality of laboratory sessions	◌	◌	◌	◌	◌	◌
Opportunity to engage in additional research activities offered by instructors	◌	◌	◌	◌	◌	◌
Student-instructor relations	◌	◌	◌	◌	◌	◌
Accessibility of instructors (student hours, e-mail, tutorial system, consultations)	◌	◌	◌	◌	◌	◌
Accessibility of student office	◌	◌	◌	◌	◌	◌
Research equipment (laboratory, computer and other technology, lecture rooms)	◌	◌	◌	◌	◌	◌
Access to literature (library, e-sources, on-line databases)	◌	◌	◌	◌	◌	◌

Q44 - What was the language of instruction in your exchange programme?

◌ English
◌ Slovenian
◌ Other:

Q45 - Please evaluate to what extent have you been satisfied with the following aspects of language accessiblity of courses at the University of Ljubljana?

	Not at all satisfied	Slightly satisfied	Mostly satisfied	Very satisfied	Completely satisfied	Does not apply to my program
Amount of courses in English language	○	○	○	○	○	○
Course materials in English language	○	○	○	○	○	○
Instructors' communication skills in English	○	○	○	○	○	○

Q46 - What is your general level of English?

○ Native speaker
○ Advanced user
○ Intermediate user
○ Elementary user

IF (3) Q46 != [1]
Q47 - What is your level of reading, listening, speaking, and writing English?

	Elementary	Intermediate	Advanced
Reading	○	○	○
Listening	○	○	○
Speaking	○	○	○
Writing	○	○	○

IF (4) Q44 != [2, 3]
Q48 - Please evaluate towhat extent has your skill of English language impacted yourperformance in oral and written participation.

	Very negatively impacted	Negatively impacted	Did not impact	Positively impacted	Very positively impacted	Does not apply to my program
Oral participation in class	○	○	○	○	○	○
Written participation (assignments, seminary papers)	○	○	○	○	○	○
Written exams	○	○	○	○	○	○

Q49 - Please evaluate how much promotional information on Slovenian culture, economy political system, history and tourism opportunities have you obtained during your exchange.

	No information	A little	A moderate amount	A lot	A great deal of information
Culture	◌	◌	◌	◌	◌
Economy	◌	◌	◌	◌	◌
Political system	◌	◌	◌	◌	◌
History	◌	◌	◌	◌	◌
Tourism opportunities	◌	◌	◌	◌	◌

Q50 - Please evaluate how much information about Slovenia you have obtained from the following sources?

	No information	A little	A moderate amount	A lot	A great deal of information
International office at the university	◌	◌	◌	◌	◌
International office at the faculty	◌	◌	◌	◌	◌
Instructors	◌	◌	◌	◌	◌
Tutors	◌	◌	◌	◌	◌
Student organizations	◌	◌	◌	◌	◌
Slovenian students	◌	◌	◌	◌	◌
Other exchange students	◌	◌	◌	◌	◌
Your consulate or embassy	◌	◌	◌	◌	◌
Other:	◌	◌	◌	◌	◌

Q51 - How often have you presented your home country to fellows in Slovenia in the following ways?

	Not at all	Rarely	Sometimes	Often	Very often
I have shared materials I brought from my home country	◌	◌	◌	◌	◌
I have presented my country through informal social events organized in Slovenia	◌	◌	◌	◌	◌
I have presented my country through conversations with other individuals	◌	◌	◌	◌	◌

Q52 - Finally, please evaluate how satisfied have you been with the following aspects of your student exchange.

	Not at all satisfied	Slightly satisfied	Moderately satisfied	Very satisfied	Completely satisfied	Not interested in this aspect
Quality of study environment	◌	◌	◌	◌	◌	◌
Student life	◌	◌	◌	◌	◌	◌
Location	◌	◌	◌	◌	◌	◌
Hospitality of locals in the university town	◌	◌	◌	◌	◌	◌
Slovenian tourist attractions	◌	◌	◌	◌	◌	◌
Amount of information I have received about Slovenia	◌	◌	◌	◌	◌	◌

Q53 - Would you recommend University of Ljubljana to someone interested in attending a student exchange programme in the same subject field as you?

◌ Definitely yes
◌ Probably yes
◌ Probably no
◌ Definitely no

Q54 - How long has your student exchange at the University of Ljubljana been in months?

[ ] months

Q55 - How long has your total stay in Slovenia been in months?

[ ] months

Q78 - Thank you very much for responding all questions to this point. We greatly appreciate it!In addition, we kindly ask you to answer 8 more questions. They are about this survey and about circumstances in which you were responding to it.

Q79 - How much did you enjoy completing this questionnaire?

◌ A great deal
◌ A lot
◌ A moderate amount
◌ A little
◌ Not at all

Q80 - How difficult was it for you to interpret the meanings of questions in this questionnaire?

- ◌ Extremely difficult
- ◌ Very difficult
- ◌ Moderately difficult
- ◌ Slightly difficult
- ◌ Not difficult at all

Q81 - How difficult was it for you to generate answers to questions in this questionnaire?

- ◌ Extremely difficult
- ◌ Very difficult
- ◌ Moderately difficult
- ◌ Slightly difficult
- ◌ Not difficult at all

Q82 - How many times did you not understand a certain word in a question?
Please give at least an approximate answer. If there were no such words, pleast write 0.

Q83 - What, if anything else, have you been doing on any electronic device while responding to this survey?
Multiple answers are possible. Please select all that apply.

- ☐ Browsing the web, reading online news and documents, reading e-books
- ☐ Texting, instant messaging or e-mailing
- ☐ Listening to music, radio, podcast or other audio content (e.g. TV in background)
- ☐ Talking on the telephone or other devices (including video chatting, e.g. Skype)
- ☐ Playing games (computer, video, web)
- ☐ Using social networks (e.g. Facebook, Twitter, etc.)
- ☐ Watching TV or video content (e.g. movies, series, news, YouTube clips)
- ☐ Working on text documents, presentations, spreadsheets, or similiar activities
- ☐ Other:
- ☐ I was not engaged in any other activities on any device.

Q84 - What, if anything else, have you been doing while responding to this survey?
Multiple answers are possible. Please select all that apply.

- ☐ Eating, drinking, or preparing a meal
- ☐ Doing household chores (cleaning, washing dishes, doing laundry)
- ☐ Taking care of other people (e.g. children)
- ☐ Talking to a person face-to-face
- ☐ Listening to a person talking (e.g. attending a lecture)
- ☐ Shopping or running errands (e.g. bank, post office)
- ☐ Walking around (e.g. taking a walk)
- ☐ Using means of transport (e.g. car, bus, train)
- ☐ Other:
- ☐ I was not engaged in any other activities.

Q85 - On what device are you responding to this survey?

- ◌ Personal computer
- ◌ Laptop computer
- ◌ Tablet computer
- ◌ Mobile phone
- ◌ Other:

Q86 - While you were responding to this survey, how many times have you been interrupted for more than 5 seconds?

Count only interruptons longer than 5 seconds.

	0 times	1 time	2 times	3 times	4 times	5-9 times	10+ times
Interruptions for activities on this device	◌	◌	◌	◌	◌	◌	◌
Interrputions for activities on other devices (e.g. TV, PC, tablet, phone, etc.)	◌	◌	◌	◌	◌	◌	◌
Other interruptions (not related to any electronic device)	◌	◌	◌	◌	◌	◌	◌

Q87 - If you wish to add anything regarding the focus of this study, your information is most welcome.

Appendix B. Screenshots

Figure B. 1: Screenshot of sketch difference

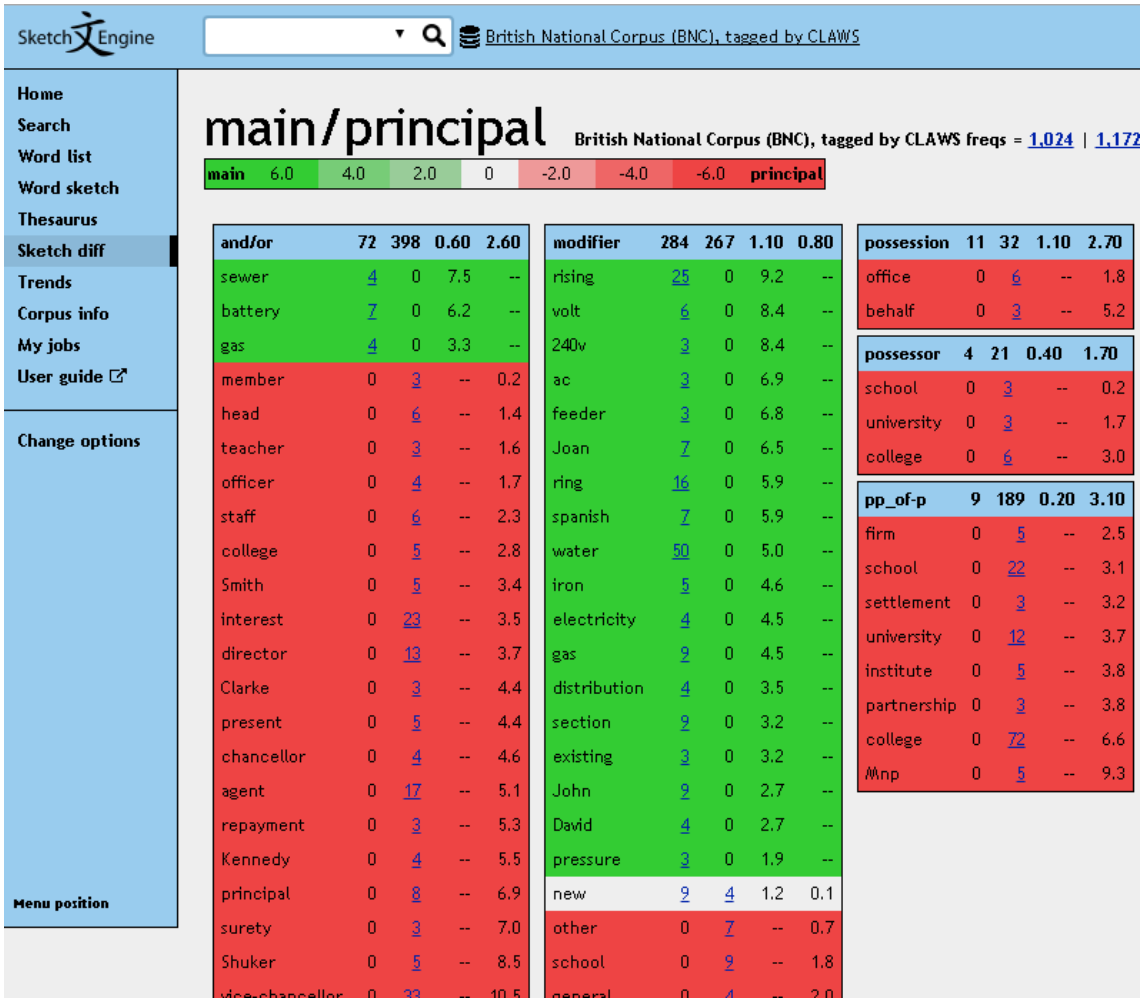


Figure B.2: WordNet screenshot for insert 'kind' example - default

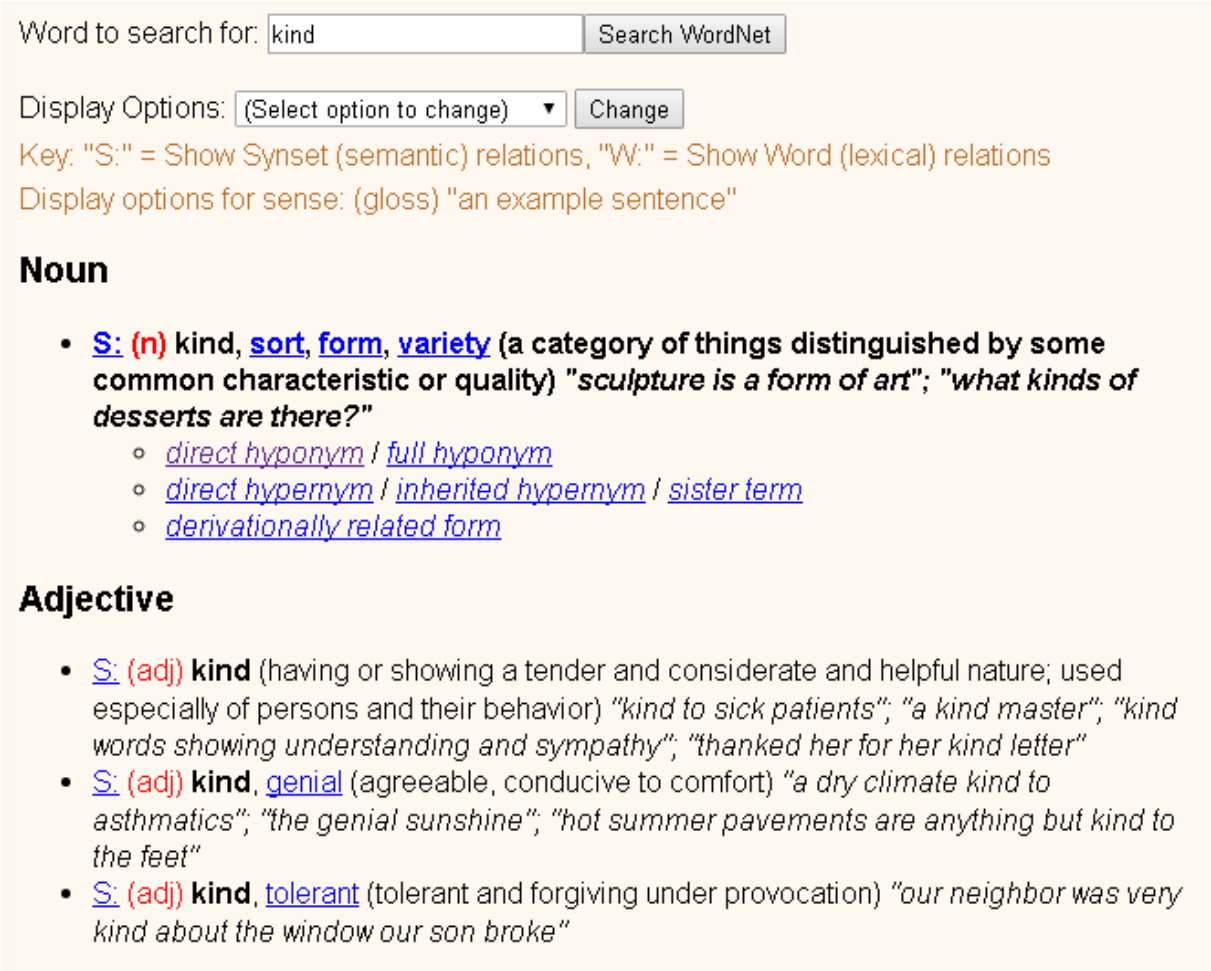


Figure B.3: WordNet screenshot for insert 'kind' example - expanded

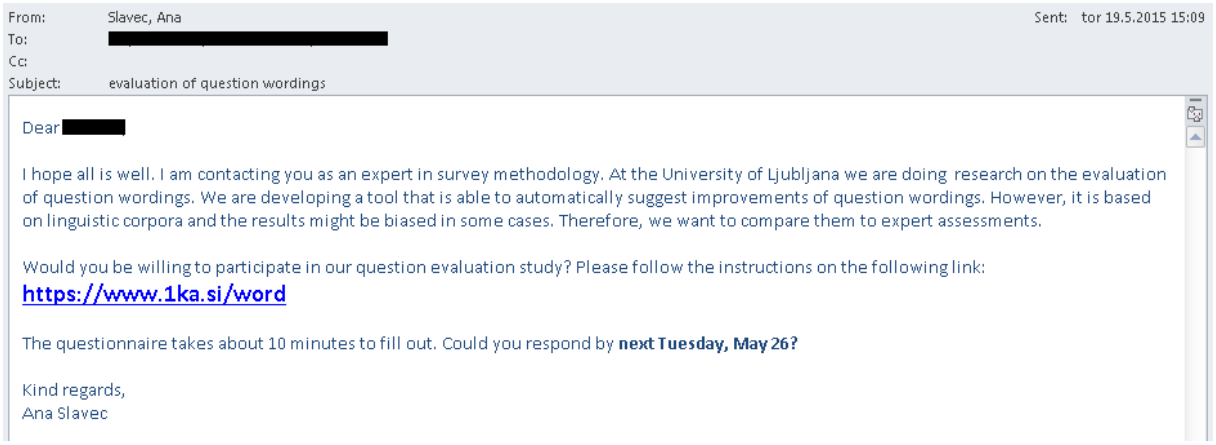
Noun

- S: (n) **kind**, sort, form, variety (a category of things distinguished by some common characteristic or quality) *"sculpture is a form of art"; "what kinds of desserts are there?"*
  - *direct hyponym* / *full hyponym*
    - S: (n) description (sort or variety) *"every description of book was there"*
    - S: (n) type (a subdivision of a particular kind of thing) *"what type of sculpture do you prefer?"*
    - S: (n) antitype (an opposite or contrasting type)
    - S: (n) art form ((architecture) a form of artistic expression (such as writing or painting or architecture))
    - S: (n) style (a particular kind (as to appearance)) *"this style of shoe is in demand"*
    - S: (n) flavor, flavour ((physics) the six kinds of quarks)
    - S: (n) color, colour ((physics) the characteristic of quarks that determines their role in the strong interaction) *"each flavor of quarks comes in three colors"*
    - S: (n) species (a specific kind of something) *"a species of molecule"; "a species of villainy"*
    - S: (n) genus (a general kind of something) *"ignore the genus communism"*
    - S: (n) brand, make (a recognizable kind) *"there's a new brand of hero in the movies now"; "what make of car is that?"*
    - S: (n) genre (a kind of literary or artistic work)
    - S: (n) like, ilk (a kind of person) *"We'll not see his like again"; "I can't tolerate people of his ilk"*
    - S: (n) manner (a kind) *"what manner of man are you?"*
    - S: (n) model (a type of product) *"his car was an old model"*
    - S: (n) stripe (a kind or category) *"businessmen of every stripe joined in opposition to the proposal"*
    - S: (n) like, the like, the likes of (a similar kind) *"dogs, foxes, and the like"; "we don't want the likes of you around here"*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) category (a general concept that marks divisions or coordinations in a conceptual scheme)
  - *derivationally related form*

Adjective

- S: (adj) **kind** (having or showing a tender and considerate and helpful nature; used especially of persons and their behavior) *"kind to sick patients"; "a kind master"; "kind words showing understanding and sympathy"; "thanked her for her kind letter"*
- S: (adj) **kind**, genial (agreeable, conducive to comfort) *"a dry climate kind to asthmatics"; "the genial sunshine"; "hot summer pavements are anything but kind to the feet"*
- S: (adj) **kind**, tolerant (tolerant and forgiving under provocation) *"our neighbor was very kind about the window our son broke"*

Figure B.4: E-mail invitation example



From: Slavec, Ana
Sent: tor 19.5.2015 15:09
To: [redacted]
Cc:
Subject: evaluation of question wordings

Dear [redacted]

I hope all is well. I am contacting you as an expert in survey methodology. At the University of Ljubljana we are doing research on the evaluation of question wordings. We are developing a tool that is able to automatically suggest improvements of question wordings. However, it is based on linguistic corpora and the results might be biased in some cases. Therefore, we want to compare them to expert assessments.

Would you be willing to participate in our question evaluation study? Please follow the instructions on the following link:
https://www.1ka.si/word

The questionnaire takes about 10 minutes to fill out. Could you respond by **next Tuesday, May 26?**

Kind regards,
Ana Slavec

Note: There was some variation in later e-mails, i.e. different dates and also the survey length estimate was adjusted after we realized that 10 minutes is not enough.

Figure B.5: Instructions for the evaluation (first page of web survey)

Dear Expert,

This is not a regular survey but an evaluation of different survey question wordings. On each page there is a question with an underlined word that can be substituted with different alternative words that are displayed on mousover.

Your task is respond to the two follow-up questions on each page:

A) Evaluate the **appropriateness** of different wording. With appropriateness we mean the wording that makes the question meaning **most clear and easy to understand** for the general population.

B) Indicate the wording would you choose and provide an explanation for your choice.

The task should require about 10 minutes in total. Please read the instructions and questions carefully. Your responses are very valuable to us.

Thank you for participating!

Razširjeni povzetek v slovenskem jeziku

Uvod

Vprašalnik je osrednji del vsake anketne raziskave in pri tem seveda želimo, da vprašalnik predstavlja veljaven in zanesljiv merilni instrument. Sestavljanje anketnih vprašanj je kompleksna naloga, ki zahteva številne konceptualne in tehnične odločitve, med drugim tudi izbiro ubeseditve vprašanja med številnimi možnimi alternativami.

Več raziskav na področju anketne metodologije je pokazalo, da so anketiranci precej občutljivi na strukturne lastnosti vprašalnikov, medtem ko je manj jasno, v kakšnem obsegu so občutljivi na razlike v ubeseditvi vprašanj (Krosnick in Fabrigar, prihodnji). Po Krosnicku in Fabrigarju so lastnosti dobre ubeseditve vprašanj univokalnost (tj. jasna osredotočenost na merjeni koncept in ločenost od drugih konceptov), enotnost pomenov (tj. ima enoten pomen za vse anketirance) ter gospodarnost z besedami (tj. uporabi se le toliko besed, kot je potrebno). Določene eksperimentalne študije so potrdile pomembnost izbire najboljše ubeseditve (npr. Kalton in drugi 1978, Duncan in Schuman 1980, Schuman in Presser 1981, Sudman in Bradburn 1983, Smith 1987, Rasinski 1989, Esposito in drugi 1991, itd.), a zaradi številčnosti možnosti lahko rečemo, da je področje še relativno neraziskano.

Raziskovanje ubeseditve vprašanj sodi v širše področje kognitivnih vidikov anketnih metod (angl. cognitive aspects in survey methods), ki temelji na psiholoških teorijah razumevanja jezika, spomina in odločanja (npr. Sirken in drugi 1999, Tourangeau in drugi 2000). V predlagani disertaciji se osredotočamo samo na razumevanje vprašanja kot stopnjo v procesu odgovarjanja, kjer je ubeseditev osrednjega pomena. Z zaznavanjem nerazumljivih vprašanj so se v anketni metodologiji ukvarjali že Graesser in drugi (1999, 2000, 2006) ter Lenzner (2011, 2012).

Medtem ko so strukturni vidiki anketnih vprašanj izčrpno obravnavani v literaturi, je problematika ubeseditve vprašanj še vedno relativno neobdelana. Zaostanek je zaznati tudi na področju razvoja računalniških modelov za detekcijo problematičnih anketnih

vprašanj. Medtem ko za strukturne značilnosti obstaja Survey Quality Predictor (SQP), računalniška aplikacija za evalvacijo anketnih vprašanj, ki temelji na metaanalizi eksperimentov za več kot 3000 vprašanj v vseh evropskih jezikih in omogoča tudi napovedi za nova vprašanja (Saris in Gallhofer 2007). Vendar je kodiranje značilnosti za nova vprašanja precej časovno potratno, poleg tega pa aplikacija ne zajema jezikovnih značilnosti ubeseditve vprašanj, ki nas zanimajo. V tej disertaciji skušamo razviti prototip aplikacije za zaznavanje nepoznanih izrazov v anketnih vprašanjih, ki bi temeljil na jezikovnih virih, kot se uporabljajo na področju računalniškega jezikoslovja, pa tudi na številnih drugih področjih, medtem ko v anketni metodologiji še niso bile uporabljene. Izjema je morda aplikacija QUAID (Graesser in drugi 2006), ki pa ima določene pomanjkljivosti, vsaj za področje nepoznanih izrazov, na katerega se osredotočamo v tej disertaciji.

Sicer se v praksi nepoznani izrazi in druge težave z ubeseditvijo anketnih vprašanj preverjajo večinoma s kvalitativnimi metodami, kot so kognitivni intervjuji in ekspertne ocene, ki temeljijo na osebni presoji in se jih izvaja v fazi predtestiranja merilnega instrumenta in za razliko od kvantitativnih pristopov ne zahtevajo že zbranih podatkov. Vendar tudi te metode lahko vzamejo precej resursov. Zato skušamo v tej disertaciji razviti nov pristop, ki bo izkoristil jezikovne vire in se bo lahko uporabljal komplementarno obstoječim metodam za predtestiranje anketnih vprašalnikov.

Namen, cilji in struktura disertacije

Glavna ideja te disertacije je razvoj nove metode za pregledovanje ubeseditev vprašanj, ki je osnovana na jezikovnih virih. Naš pristop gradi na Krosnickovi ideji, da bi se anketna vprašanja za anketirance poenostavilo z uporabo posebnih računalniških programov, ki zaznajo nerazumljive in dvoumne besede, ter predlagajo sopomenke, ki so razumljivejše in jasnejše. Pri tem nadgrajujemo delo Graesserja in Leznerja, ki sta že razvila določene postopke za zaznavanje težav z razumljivostjo vprašanj, vendar ne ponujata predlogov sprememb.

Naloga zagovarja tezo, da je možno razviti postopek za predtestiranje anketnih vprašalnikov, ki temelji na tekstovnih korpusih in leksikalnih bazah, s katerim je možno

učinkovito zaznati težko razumljive ubeseditve in predlagati alternativne ubeseditve. Pri tem smo odgovarjali na pet raziskovalnih vprašanj:

1. Kako izbrati in kombinirati različne jezikovne vire tako, da bodo razvijalci anketnih vprašalnikov lahko zaznali ubeseditve vprašanj z nizko frekvenco v korpusih?
2. Ali imajo strokovnjaki za anketne metode ubeseditve z višjo frekvenco v korpusih za primernejše kot tiste ubeseditve, ki imajo nižjo frekvenco v korpusu, a enak pomen?
3. Ali udeleženci v kognitivnih intervjujih ubeseditve z višjo frekvenco v korpusu razumejo bolje kot ubeseditve z nižjo frekvenco v korpusu?
4. Ali je kakovost odgovorov boljša v anketah, ki uporabljajo ubeseditve z visoko frekvenco v korpusih namesto ubeseditev z nizko frekvenco?
5. Kako jezikovne vire integrirati v postopke razvoja anketnega vprašalnika?

Po uvodni definiciji problema v prvem poglavju je v drugem poglavju disertacije predstavljeno teoretsko ozadje. V tretjem poglavju predstavimo pilotno študijo na primeru angleške in slovenske različice vprašalnika za študente na mednarodni izmenjavi na Univerzi v Ljubljani, kjer na podlagi jezikovnih virov razvijemo dve različici obeh vprašalnikov in ju primerjamo v eksperimentu z deljenim vzorcem. V četrtem poglavju pristop na podlagi jezikovnih virov primerjamo z ekspertnimi ocenami, in sicer na dveh študijah primera, vprašalniku Wageindicator o plačah in delovnih pogojih ter izbranih vprašanjih PEW raziskovalnega centra na temo terorizma. Slednja nadalje proučujemo tudi v petem vprašanju, kjer jih uporabimo v kognitivnih intervjujih, ter tudi v šestem poglavju, kjer na podlagi jezikovnih virov razvijemo še tri različice teh vprašanj in jih primerjamo v eksperimentu z deljenim vzorcem. V sedmem poglavju predstavimo prototip aplikacije za zaznavanje nepoznanih in dvoumnih izrazov v anketnih vprašanjih. Osmo poglavje pa je zaključni sklep o smotrnosti pristopa na podlagi jezikovnih virov kot metode predtestiranja in evalvacije anketnih vprašalnikov.

Teoretsko ozadje

Najprej uvedemo osnovne pojme korpusnega jezikoslovja in leksikalne semantike ter predstavimo jezikovne vire, ki jih bomo uporabili za analizo jezikovnih lastnosti vprašanj, predvsem pogostosti besede v izbranih besedilih. Nato predstavimo že omenjeno aplikacijo QUAID za zaznavanje potencialno težavnih anketnih vprašanj ter izpostavimo njene pomanjkljivosti. Sledi podrobnejši opis ustaljenih pristopov za predtestiranje anketnih vprašanj. To so predvsem kvalitativne metode, na primer kognitivni intervjuji in ekspertne ocene, pa tudi kvantitativni pristopi, kjer opazujemo različne indikatorje kakovosti odgovorov, ki jih lahko v primeru eksperimentov z deljenim vzorcem tudi primerjamo. V tem poglavju tudi predstavimo nekaj primerov takih raziskav, ki so primerjale različne verzije istih vprašalnikov, predvsem tiste, ki so se osredotočale na razlike v razumljivosti besedišča.

Korpusno jezikoslovje in leksikalna semantika

Korpusno jezikoslovje je raziskovanje naravnega jezika na podlagi obsežnega empiričnega vzorca besedil iz vsakdanje jezikovne rabe, ki lahko obsega vse od knjig in časopisov. Besedilni korpusi so torej velike zbirke besedil v naravnem okolju, ki se lahko uporabijo kot mera pogostosti uporabe določene besede ali besedne zveze v jeziku. Domnevamo namreč, da višja, kot je frekvenca v korpusu, bolj je beseda poznana splošni populaciji.

V disertaciji uporabljamo tri besedilne korpuse za angleški jezik in en korpus za slovenski jezik. Za angleščino uporabimo dva največja načrtovana korpusa, British National Corpus (BNC) (Burnard 1995; Leech in drugi 2001) za britansko angleščino in Corpus of Contemporary American English (COCA) (Davies 2010) za ameriško angleščino. Njuna prednost je, da sta nastala načrtovano in vsebujeta tudi metapodatke, tj. podatke o žanru, vrsti besedila, letu itd. Na podlagi teh podatkov je možno korpus tudi uravnotežiti. Tretji angleški korpus pa je enTenTen, ki je še večji kot BNC in COCA, vendar ni načrtovan. Nastal je namreč na podlagi avtomatskega zajemanja besedil s spleta (angl. web crawling), ki so nato še prečiščena (Jakubiček in drugi 2013).

Za slovenski jezik pa uporabimo korpus KRES, ki je uravnotežen podvzorec korpusa pisne slovenščine Gigafida (Logar Berginc in drugi 2012).

Glavna funkcija vseh naštetih jezikovnih korpusov je konkordančnik, programski vmesnik, ki omogoča iskanje po korpusu in izpis zadetkov, ki ustrezajo določenemu iskalnemu nizu. Med drugim lahko tudi preverimo, kako pogosto se določena beseda ali besedna zveza uporabi v korpusu. Ta frekvenca je premosorazmerna z velikostjo korpusa, zato je treba frekvence različnih korpusov normalizirati, da jih lahko primerjamo. Obstajajo pa tudi naprednejša orodja za uporabljanje s korpusi, na primer Sketch Engine, s katerim je možno analizirati vse zgoraj omenjene korpuse. Poleg konkordance omogoča tudi analizo kolokacij in besedne skice.

Ker nas v disertaciji zanimajo predvsem frekvence, velja omeniti Zipfov zakon, ki pravi, da se pogostost porazdeljuje po funkciji ena deljeno z n na kvadrat, kar pomeni, da je najpogostejša beseda približno dvakrat pogostejša od druge najpogostejše besede, ta pa dvakrat pogostejša od četrte najpogostejše besede, in tako dalje. Podoben zakon obstaja tudi za porazdelitev števila pomenov besed v leksikonih, ki pravi, da imajo pogostejše besede tudi večje število pomenov.

Poleg jezikovnih korpusov v disertaciji uporabljamo še leksikalne baze, in sicer kot vir sopomenk in drugih alternativnih ubeseditev, s katerimi lahko potencialno problematične besede zamenjamo s pogostejšimi alternativami, po možnosti z enakim pomenom. Taka baza je na primer WordNet (Miller 1995), ki je poskus organizacije leksikalnih informacij glede na pomene besed namesto glede na obliko, kot je običajno praksa v slovarjih. Besedi sta v WordNetu sopomenki, če si delita vsaj en pomen, v katerem je besedi mogoče zamenjati, ne da bi s tem spremenili pomen stavka. Take besede so skupaj v sinsetu. Obstajajo tudi wordneti v drugih jezikih, na primer slovenski sloWNet (Fišer 2009). Wordneti so imeli veliko aplikacij na raznih področjih, medtem ko za izboljševanje anketnih vprašanj še niso bili uporabljeni. Poleg sopomenk lahko v wordnetih iščemo tudi nadpomenke, podpomenke in podobne izraze. Ena beseda je drugi nadpomenka, ko jo pojmovno ali vsebinsko vsebuje, medtem ko je podpomenka beseda, ki je bolj specifično določena od že dane besede ali besedne zveze in ima ožji pomen od svoje nadpomenke.

Jezikovni viri v procesu načrtovanja anketnega vprašalnika

Čeprav so tako jezikovni korpusi in leksikalne baze prosto dostopne za akademsko rabo in bi lahko bile uporabljene za izračun indikatorjev redkosti besede ter njene nejasnosti, ostajajo v anketni metodologiji neraziskane. Izjema je morda le aplikacija QUAID, ki se osredotoča specifično na zaznavanje potencialno težavnih anketnih vprašanj z lingvističnega vidika (Graesser in drugi 1999, 2006), vendar je naša izkušnja z aplikacijo, da daje precej lažno pozitivnih rezultatov. Poleg tega metoda ne daje predlogov popravkov in izboljšav vprašanj, ampak jih mora uporabnik poiskati sam.

Kvalitativne metode za evalvacijo anketnih vprašanj

Anketne vprašalnike lahko predtestiramo z različnimi kvalitativnimi metodami, med katerimi sta dve izmed najbolj razširjenih kognitivni intervjuji in ekspertne ocene.

Kognitivni intervju je terenska metoda raziskovanja, s katero zbiramo podatke o udeležencevih kognitivnih procesih. Metoda omogoča izjemen fokus in z njo lahko raziskovalec zbere podatke o procesu odgovarjanja ter identificira probleme, ki jih drugače ne bi opazil (Willis 1999; Mohorko in Hlebec 2013). Med drugim se jih lahko uporabi tudi za zaznavanje težav z razumevanjem vprašanj. Uporabljajo se različne tehnike, od katerih v disertaciji uporabimo dve: parafraziranje in definicije. Parafraziranje pomeni, da respondenta prosimo, da na vprašanje odgovori z lastnimi besedami, definicije pa, da definira določen izraz v vprašanju.

Še ena metoda za evalvacijo so ekspertne ocene, ki izkoriščajo znanje strokovnjakov na področju oblikovanja anketnih vprašalnikov in je lahko zelo učinkovita, zlasti v začetnih fazah razvoja vprašalnika (Lessler in Forsythe 1996; Akkerboom in Dehue 1997). Ekspertne ocene so bile že uporabljene tudi za zaznavanje težav z jezikovno strukturo (Holbrook in drugi 2007). Čeprav je metoda učinkovita za detekcijo težavnih vprašanj, je omejitev to, da je manj zanesljiva (Willis 1999; Cerar in drugi 2001; Olson 2010; Saris 2012; Yan in drugi 2012) in rezultati lahko zelo variirajo od respondenta do respondenta (Olson 2010). Graesser (2000; 2006) in Olson (2010) sta ekspertne ocene kot metodo za zaznavanje težav z razumevanjem vprašalnika primerjala z aplikacijo

QUAID. Čeprav je bilo veliko ujemanja, je bilo veliko tudi lažno pozitivnih in drugih neujemanj.

Tako ekspertne ocene kot kognitivni intervjuji so resursno potratni, zato bi bila dopolnilna metoda za zaznavanje tovrstnih težav na podlagi jezikovnih baz zelo dobrodošla.

Indikatorji kakovosti odgovorov v anketnih raziskavah

Poleg kvalitativnih metod predtestiranja poznamo tudi kvantitativne, ki pa že zahtevajo pilotno raziskavo ali pa celo študijo na večjem vzorcu anketirancev. Izvede se jih torej ob zbiranju podatkov in omogočajo izboljšanje šele v naslednji izvedbi raziskave. Metoda, ki jo v tej disertaciji uporabljamo so eksperimenti z deljenim vzorcem (Rug 1941, Cantril 1943), kjer je vsaka enota naključno uvrščena v eno izmed dveh (ali več) skupin.

Kar primerjamo med skupinami, pa je izbor različnih indikatorjev kakovosti anketnih podatkov. Težko razumljiva vprašanja na kakovost odgovorov vplivajo na različne načine. Najprej, četudi anketiranec razume vprašanja, zahtevnost razumevanja poveča njegovo obremenitev (angl. response burden) (Bradburn 1978), kar se odraža v daljših časih odgovora. Drugič, zaradi težavnosti odgovarjajoči lahko dela kognitivne bližnjice pri odločanju o odgovorih in zato ne odgovarja optimalno. Pojav se imenuje zadostovanje (angl. satisficing) (Krosnick 1991, 1996) in se odraža v nerazlikovanju odgovorov, pogostejšemu izbiranju vedno prve ali srednje vrednosti na lestvici, izbiranju odgovora »ne vem« ipd. Tretjič, anketiranec se lahko odzove tudi z neodgovorom na določena vprašanja ali celo tako, da zapusti anketo pred njenim koncem. Četrtič, anketiranec, ki narobe razume vprašanja, lahko poda odgovor, ki ne ustreza resničnemu stanju. Za vse naštete pristope obstajajo določene kvantitativne mere.

Poleg tega lahko v vprašalniku izmerimo tudi subjektivno zaznavo zahtevnosti vprašanj, in sicer tako, da na koncu vprašalnika postavimo nekaj dodatnih vprašanj (Hedlin 2005).

Prejšnje študije z deljenim vzorcem

V literaturi je veliko primerov uporabe eksperimentov z deljenim vzorcem za primerjavo različnih ubeseditev vprašalnika. Med bolj znanimi eksperimenti so na primer študija na vprašanju o zanimanju za religijo (Duncan in Schuman 1980) ter več eksperimentov na ameriški General Social Survey (Rasinski 1989; Smith 1987).

V tej disertaciji pa so nas zanimale predvsem tiste študije, ki so v primerjavo vključile koncept pogostosti besed. Čeprav ne omenja frekvenc, je taka na primer študija Blasiusa in Friederichsa (2009), ki sta variirala ubeseditev sedmih trditev v matričnem vprašanju, pri čemer sta uporabljala vsakdanje besedišče in bolj elaboriran jezik. Za tri od postavk se je pokazala razlika v porazdelitvi odgovor.

Še bližje temi te disertaciji pa je več Lenznerjevih raziskav, ki je v svojem delu tudi uporabljal nepoznane izraze kot eno od psiholingvističnih determinant zahtevnosti vprašanja (Lenzner in drugi 2010; Lenzner 2011; Lenzner in drugi 2011; Lenzner 2012). V eksperimentu z deljenim vzorcem, kjer je ena različica imela pogostejši, druga pa redkejši izraz, je primerjal čas odgovarjanja, prekinitve in stopnjo zadostovanja. Ugotovil je, da ima verzija z redkejšimi izrazi daljši čas odgovarjanja, medtem ko učinka na stopnjo prekinitev in zadostovanje ni mogel potrditi (Lenzner in drugi 2010). V nadaljevanju je anketna vprašanja preveril še z metodo spremljanja gibanja oči (angl. eye-tracking) in potrdil, da se pogled respondentov dlje časa zadrži pri postavkah z nižjimi frekvencami (Lenzner in drugi 2011). Nato je izvedel še en eksperiment z deljenim vzorcem, tokrat na večjem vzorcu in ugotovil, da izboljšanje razumljivosti zmanjša obseg nevsebinskih odgovorov, kot je na primer »ne vem« ipd. (Lenzner 2010).

Pilotna študija

Najprej smo izvedli preliminarno pilotno študijo na dveh anketnih vprašalnikih za študente na mednarodni izmenjavi na Univerzi v Ljubljani, pri čemer je bil eden v angleškem (za prihajajoči študente) in eden v slovenskem jeziku (za odhajajoče študente). Oba vprašalnika smo pregledali z jezikovnimi viri in razvili dve različici obeh vprašalnikov, eno z nizkimi frekvencami ubeseditev in drugo z visokimi frekvencami ubeseditev. Skupaj sta se obe angleški različici razlikovali v 23 ubeseditvah, slovenski pa v približno 40 ubeseditvah.

Obe različici smo nato primerjali v dveh eksperimentih z deljenimi vzorcema, kjer je bila polovica vzorca naključno dodeljena kontrolni skupni, ki je odgovarjala na različico z nizkimi frekvencami, in drugi polovici, ki je bila dodeljena eksperimentalni skupini, ki je odgovarjala na različico z višjimi frekvencami. Vabilo za sodelovanje je bilo poslano 1147 tujim študentom in 917 slovenskim, v raziskavi pa je na koncu sodelovalo 230 tujih (20% stopnja odgovora) in 205 (25% stopnja odgovora) študentov.

Rezultati kažejo, da je bilo manj prekinitev odgovarjanja v dveh različicah z visokimi frekvencami. Poleg tega so anketiranci v slovenski različici s pogostejšimi izrazi poročali nižje število besed, ki niso bile razumljene. Rezultati se v določenih vidikih ujemajo s prejšnjimi študijami o učinku uporabe besed z višjo frekvenco, vendar so tudi določena razhajanja. Tako kot Lenzner (2010; 2012) nismo v angleški in slovenski različici opazili nobene razlike v stopnji neodgovora spremenljivke ter v zadostovanju. Za razliko od Lenznerja nismo potrdili razlike v času odgovora, medtem ko nam je uspelo pokazati učinek na stopnjo prekinitev, česar Lenzner ni uspel potrditi.

Vendar je vzorec te študije zelo majhen, študentska populacija preveč specifična in eksperimentalni načrt zelo osnoven. Še ena omejitev je uporaba posameznih besed namesto daljših nizov besed, kar smo nadgradili v sledečih empiričnih študijah.

Primerjava ekspertnih ocen in jezikovnih virov

Druga empirična študija je bila na vzorcu in je vključevala primerjavo pristopa na podlagi besedilnih korpusov z ekspertnimi ocenami za zaznavanje nepoznanih izrazov. Dva niza anketnih vprašanj sta bila izbrana kot študiji primera: prva je bila izbor osmih anketnih vprašanj (sedem različnih ubeseditev) iz vprašalnika WageIndicator o plačah in delovnih pogojih, druga pa je bila izbor osmih vprašanj (12 postavk in 12 različnih ubeseditev) iz baze anketnih vprašanj PEW. Oba vprašalnika smo evalvirali na podlagi jezikovnih korpusov, alternativne izraze pa smo poiskali v leksikalni bazi WordNet, in sicer smo za vsako postavko izbrali nekaj besed, ki so jih potem evalvirali eksperti. Eksperte smo prosili, naj ocenijo primernost različnih ubeseditev, označijo, katere bi izbrali, in komentirajo svoje odgovore. K sodelovanju smo povabili 132 ekspertov, skupno pa jih je sodelovalo 81 ekspertov. Od tega jih je 17 ocenjevalo obe različici, ostali pa samo eno od dveh.

Rezultati so pokazali, da se evalvacije ekspertov in besedilni korpusi ujemajo za več kot polovico postavk, in sicer za pet od devetih primerov v študiji Wageindicator ter v sedmih od enajstih primerov v študiji PEW. V preostalih štirih primerih v prvi in štirih v drugi študiji pa so bile razlike, kar lahko delno razložimo s tem, da besede niso imele povsem enakega pomena in zato niso zamenljive v tej situaciji – z drugimi besedami, niso sinonimi. Konkretno so bila pri prvi študiji neujemanja med metodama za pridevnika 'wholly' in 'completely', glagola 'set' in 'made', prislova 'enough' in 'sufficiently' ter prislova 'adequate' in 'enough'. Pri drugi študiji pa so bile razlike med metodama za štiri pare glagolov, in sicer 'worried' in 'concerned', 'restrict' in 'limiting', 'gathering' in 'collecting' ter 'promote' in 'encourage'. Poleg tega so bile v nekaterih primerih zaznane razlike med eksperti, katerih materni jezik je angleščina, in tistimi, ki jim ni.

Rezultati se ujemajo s študijama Graesserja in drugih (2000) ter Olson (2010), ki so ekspertne ocene primerjali z rezultati na podlagi aplikacije QUAID. Zaključimo lahko torej, da polavtomatski pristop na podlagi korpusov lahko zamenja resursno zahtevne ekspertne evalvacije. Vendar je za končno odločitev opraviti še dodatne analize za posamezne pare besed.

Kognitivni intervjuji

Tretjo empirično študijo sestavlja 122 spletnih kognitivnih intervjujev, kjer smo udeležence vprašali, naj bodisi definirajo določeno ubeseditev v anketnem vprašanju, bodisi parafrazirajo celotno vprašanje. V celoti smo evalvirali 13 postavk, vse iz zgoraj omenjenega niza vprašanj PEW. Udeležence smo rekrutirali z uporabo platforme Prolific Academic za množično sodelovanje ('crowdsourcing'), za sodelovanje pa so bili plačani 1,25 funta. Rekrutacija je potekala v dveh valovih: v prvem smo rekrutirali splošen vzorec 60 udeležencev, v drugem valu pa smo rekrutirali še 63 oseb, katerih materni jezik ni angleščina. Študija je bila osnovana na eksperimentu z deljenim vzorcem, saj je bila polovica sodelujočih (v obeh skupinah) naključno razvrščena v različico z izvirnimi vprašanji PEW, polovica pa v različico z izboljšanimi (sedem primerov) ali poslabšanimi (šest primerov) vprašanj.

Rezultati so pokazali visoko ujemanje (tj. veliko število podobnih odgovorov med vzorcema) za pridevnika 'careful' in 'cautious', samostalnika 'threat' in 'menace' ter samostalnik 'ransom' in pridevnik 'demanded (money)'. Srednja stopnja ujemanja je bila zaznana za pridevnik 'sympathetic' in glagol 'support', glagola 'collecting' in 'gathering', pridevnika 'prone' in 'inclined' ter samostalnika 'chances' in 'risk'. Slabo ujemanje pa je bilo med pridevnikoma 'worried' in 'apprehensive', pridevnikoma 'justified' in 'vindicated', glagoloma 'restricting' in 'limiting', glagoloma 'encourage' in 'promote' ter pridevnikoma 'concerned' in 'preoccupied'.

Ugotovili smo, kot pričakovano, da v primeru, ko uporabimo besedo z nizko frekvenco v korpusu, to besedo udeleženci definirajo z njeno bolj frekventno alternativo. Še ena ugotovitev je, da smo v primeru ubeseditev z višjo frekvenco skupno našteli višje število različnih definicij in parafraz, v primerjavi z njihovimi nizko frekventnimi alternativami. V nekaterih primerih smo to pojasnili z višjim številom pomenov (v bazi WordNet), kar nakazuje večjo dvoumnost teh izrazov. Poleg tega smo ugotovili tudi določene razlike med tistimi, ki jim je angleščina materni jezik, in med tistimi, ki jim ni.

Glavna študija

Četrta in glavna empirična študija je bila eksperiment z deljenim vzorcem, kjer smo primerjali štiri verzije istega vprašalnika PEW: izvorno, izboljšano (11 zamenjav z bolj frekventnimi ubeseditvami), slabšo (16 zamenjav z manj frekventnimi ubeseditvami) in najslabšo (34 zamenjav z manj frekventnimi ubeseditvami). Različice smo poimenovali različica -2 (najslabša), -1 (slaba), 0 (izvorna) in 1 (izboljšana).

Eksperiment je bil izveden na ameriškem neverjetnostnem panelu Survey Monkey Audience. Udeleženci so za sodelovanje nagrajeni tako, da izberejo dobrodelno organizacijo, ki naj ji podjetje nakaže nagrado. Točno število vabil k sodelovanju v anketi, ki jih je poslal Survey Monkey, ni znano, je pa anketo začelo reševati 2966 oseb, od tega pa jih je 2557 doseglo zadnjo stran (86 %). Glavne ugotovitve pa so naslednje:

- Stopnja prekinitev odgovarjanja je bila najvišja v najslabši verziji (-2), in sicer 17 %, medtem ko je v ostalih treh verzijah le okrog 13 %. Do največ prekinitev je prišlo pri petem vprašanju, ki je matričnega formata, in tudi na tisti točki ima najslabša verzija približno štiri odstotne točke več kot ostale tri.
- V povprečju je bilo trajanje vprašalnika okrog 5 min, vendar zahtevnejša je bila različici, dlje časa so respondenti odgovarjali na vprašalnik. Verzija -2 ima statistično značilno daljši čas odgovarjanja.
- Glede deleža tistih, ki so izbrali odgovor »ne vem«, ni statistično značilnih razlik, razen pri enem od primerov, kjer sta imeli različici, ki sta uporabili besedo 'vindicated', tri odstotne točke več odgovorov »ne vem« kot različici, ki sta uporabili besedo 'justified'.
- Tudi pri težnji po strinjanju, merjeni kot delež tistih, ki so na določeno trditev v matričnem vprašanju odgovorili »se strinjam« ali pa celo »zelo se strinjam«, ni večjih razlik med skupinami, razen za dve vprašanji.
- Pri subjektivnih ocenah zahtevnosti vprašalnika je opazna razlika, da so tistih v slabših verzijah vprašalnika, tega ocenili kot zahtevnejšega in tudi poročali o višjem številu besed, ki jih niso razumeli. Rezultate smo testirali tudi s kontrolnimi spremenljivkami spol, izobrazba in materni jezik. Moški, manj izobraženi in tisti, ki jim angleščina ni materni jezik,

so namreč vprašalnik ocenili kot zahtevnejšega kot ženske, bolj izobraženi in tisti, ki jim je angleščina materni jezik. Pri preverjanju, kako to vpliva na odnos med zahtevnostjo in različico, je pri spolu ta učinek precej manjši, glede na izobrazbo in materni jezik pa se povezava pokaže samo pri izobraženih in tistih, ki jim je angleščina materni jezik. Po drugi strani pa tega učinka ni med neizobraženimi in tistimi, ki jim angleščina ni materinščina. Verjetno zato, ker je vprašalnik za njih že v osnovi zahtevnejši.

- Podrobnejše proučevanje korpusnih frekvenc je pokazalo, da v primerih, kjer smo našli statistično značilne razlike v porazdelitvi odgovorov in/ali v kakovosti odgovarjanja med različicami vprašalnika, dejavnik ni toliko razmerje sprememb, ampak nizke frekvence izraza, ki smo ga izboljševali.

Eksperiment je potrdil, da poznanost izraza, kot jo merimo s frekvencami v korpusih, lahko vpliva na različne vidike kakovosti anketnih podatkov, zlasti na prekinitve odgovarjanja in subjektivne ocene težavnosti odgovarjanja; zaznali pa smo tudi daljši čas odgovarjanja in za nekatere postavke tudi več odgovorov »ne vem« ter večjo težnjo k strinjanju z odgovori. Vendar so učinki večinoma majhni. Zdi se, da le manjše število sprememb ne povzroči velike spremembe pri večini indikatorjev kakovosti odgovarjanja.

Zaključek

Anketno zbiranje podatkov je prevladujoča metoda v kvantitativnem družboslovnem raziskovanju in pisanje dobrih anketnih vprašanj je pomemben del zagotavljanja visoke kakovosti zbranih podatkov. Pri tem je izbira najbolj optimalne ubeseditve med več alternativami eno od bolj zapletenih vprašanj v postopku razvoja vprašalnika. Obstaja več metod za evalvacijo anketnih vprašanj, vendar so pogosto zelo kompleksne in porabijo preveč resursov. Eden od pristopov, kako narediti izboljšave preprostejše in dostopnejše, je osnovan na uporabi jezikovnih virov, ki so bili na področju anketne metodologije do sedaj le redko uporabljeni. V disertaciji smo proučevali, kako lahko jezikovne vire uporabimo za izboljšanje anketnih vprašanj.

Rezultati so potrdili, da je na osnovi besedilnih korpusov, leksikalnih baz in slovarjev možno razviti postopek, na podlagi katerega lahko učinkovito zaznavamo problematične ubeseditve anketnih vprašanj in predlagamo alternativne. Obstaja povezava med stopnjo ujemanja različnih ubeseditev in učinki na proces odgovarjanja. Poleg tega rezultati kažejo, da je v večini primerov pristop na osnovi korpusov primerljiv z ekspertnimi ocenami in kognitivnimi intervjuji. Vendar je pomembno, da pri tem upoštevamo specifičnost zasnove različnih korpusov in se ne omejimo na evalvacijo le posameznih besed, ampak da preverimo tudi daljše besedilne nize.

Omejitve raziskave in smeri prihodnjega raziskovanja

Če povzamemo, so ključne omejitve empiričnih študij v tej disertaciji naslednje:

- Čeprav so bili primeri za empirične študije zelo skrbno izbrani in smo skušali učinke različnih ubeseditev čim bolje izolirati od ostalih, lahko specifike izbranih primerov vplivajo na naravo zaključkov. S tem je povezan tudi pomislek, ali so učinki realistični ali so produkt pretiravanja. V številnih primerih namreč nismo izboljševali anketnih vprašanj, ampak jih poslabševali. Vprašamo se lahko, ali bi kdo naravno tako ubesedil vprašanja, kot so bila v najslabših različicah. Po drugi strani pa so izbrani primeri morda predobri in bi bilo v primeru izbire slabših vprašanj več prostora za izboljšave.

- Izboljšave vprašanj bi morali prilagoditi proučevanim populacijam. Za pilotno študijo, v kateri so sodelovali študentje, bi na primer lahko uporabili bolj specifične korpuse, ki bi bili na osnovani na besedilih, ki so bolj poznana tej populaciji. Tudi v primerih na splošni populaciji bi lahko ubeseditve optimizirali tako, da bi jih prilagajali različnim stopnjam izobrazbe in drugih značilnostim respondentov, vendar to v praksi ni preprosto izvedljivo. Zato je naše priporočilo za ankete na splošni populaciji, da se poskuša delati izboljšave, ki koristijo najšibkejšim članom, medtem ko naj se prilagajanje uporablja le za vprašalnike za specifične populacije (npr. vprašalnik, namenjen odvetnikom).
- V glavnem eksperimentu je bilo nekaj težav s parapodatki, zato nismo mogli pridobiti časov odgovarjanja po posameznih straneh. Zato ni bilo možno opraviti naprednejših analiz zakasnitev odgovarjanja.
- Omejitev je tudi to, da so bile v približno četrtini primerov pripadajoče korpusne frekvence (za daljše besedne nize) prenizke.

V prihodnosti bi se bilo treba osredotočiti predvsem na model, ki bi za sopomenke lahko predvidel učinek spremembe v korpusnih frekvencah, na proces odgovarjanja, na anketno vprašanje in na indikatorje kakovosti podatkov. Na tej podlagi bi lahko določili kritični nivo sprememb ubeseditev, ki lahko ogrozijo kakovost anketnih podatkov. Poleg tega so za odkrivanje ključnih faktorjev v tej kompleksni zadevi potrebne sistematične metaanalitične študije raznih sekundarnih podatkov. Poudariti gre tudi potencial, da se ta pristop vključi v obstoječa programska orodja za spletno anketiranje. Zato predlagamo naslednje smeri prihodnjih raziskav:

- Uporaba pristopa na študijah primera, kjer so bile že uporabljene druge metode testiranje vprašalnikov. Obstajajo namreč poročila o predtestiranju tako za ekspertne ocene kot za kognitivne intervjuje, kar bi omogočilo dodatne primerjave med različnimi metodami, kjer bi se lahko osredotočili na prečna preverjanja učinkovitosti različnih metod predtestiranja za zaznavanje težav z ubeseditvijo vprašanja.
- Boljši izkoristek funkcionalnosti, ki jih ponujajo jezikovna orodja, na primer Sketch Engine. Najprej bi bilo treba izkoristiti korpusne

metapodatke in rezultate prilagoditi različnim žanrom, virom, času publikacije in tako dalje. Specifična metrika, ki bi lahko bila uporabna, je razpršenost besede po različnih virih. Drugič, prihodnje raziskave bi morale iti preko preprostih konkordanc in izračunati tudi kolokacije besed v določenem kontekstu. Do neke mere je to že možno z uporabo besednih skic, vendar format iskalnika zaenkrat ne omogoča izbire specifičnega konteksta.

- Razširitev diagnostike na druge indikatorje razumljivosti vprašanja, kot so na primer razni indikatorji razumljivosti.

Integracija v obstoječa programska orodja za spletno anketiranje

Ena od motivacij za to študijo je bila tudi ideja razviti prototip orodja oziroma programa, ki bi omogočil označevanje nepoznanih besed in predlagal izboljšave. Razvili smo pilotno aplikacijo (za angleški in slovenski jezik), ki na podlagi frekvenc v korpusu označi nepoznane izraze ter predlaga sopomenke in druge alternativne izraze na podlagi WordNeta. Aplikacija je vključena v programsko orodje za spletno anketiranje 1KA (www.1ka.si), in sicer v navigaciji Testiranje – Jezikovni pregled, kjer jo lahko razvijalci anketnih vprašalnikov uporabljajo za preverjanje pogostosti različnih ubeseditev in za iskanje alternativnih izrazov.

Trenutno aplikacija deluje samo za posamezne besede, vendar načrtujemo razširitev na daljše besedne nize. Kvalitativne študije narejene v okviru te disertacije so namreč pokazale, da je kontekst, v katerem se pojavi beseda, precej pomembnejši kot samo frekvenca posamezne besede. Zato bi bilo treba aplikacijo nadalje razviti in poleg konkordance vključiti tudi kolokacije in besedne skice. Poleg tega bi bilo smiselno v proces testiranja vključiti še druge mere kompleksnosti anketnega vprašalnika, kot so koeficienti berljivosti in podobno, ter raziskati tudi druge vidike kompleksnosti ubeseditve anketnega vprašanja, na primer abstraktnost besedišča, dvojna zanikanja, vodilna vprašanja in druge oblike dvoumnih vprašanj. Vendar za te indikatorje zaenkrat še ne obstajajo preproste metode, ki bi zaznale težave in predlagale alternativne ubeseditve.

Izvirni prispevek

Izvirni prispevek disertacije področju anketne metodologije je v novem postopku za izboljševanje anketnih vprašanj, ki povezuje statistične algoritme in jezikovne vire z najnovejšimi izsledki na področju psiholingvistike in računalniške leksikografije. S teoretičnega vidika disertacija pripeva k raziskavam, povezanim s kognitivnimi vidiki razvoja vprašalnika, ter h konceptualizaciji in boljšemu razumevanju vloge informacijsko-komunikacijskih tehnologij v anketnem procesu. S praktičnega vidika pa disertacija prispeva k zasnovi postopka, ki se bo lahko uporabljal kot komplementarna metoda predtestiranja anketnega vprašalnika.