Compiling a Corpus-Based List of Words Commonly Mispronounced

Magdalena Zając

Abstract The inspiration for the paper was professor Sobkowiak's list of Words Commonly Mispronounced (Sobkowiak, 2001), a collection of over six hundred pronunciation errors that are habitually made by Polish learners of English. The paper explores the ways in which lists such as Words Commonly Mispronounced could be "upgraded" using corpus linguistic tools. The paper describes the results obtained in a previous study (Zając & Pęzik, 2012), whose aim was to compile a corpus-based index of frequent mispronunciations in the speech of Polish learners of English and which used data from the spoken component of the Polish Learner English Corpus PLEC. The paper discusses the list obtained by Zając and Pęzik, describes and evaluates the process of creating the list, and compares the corpus-based index with Words Commonly Mispronounced. The difficulties related to the compilation of lists of common mispronunciations (both corpus-based and "traditional") are also examined. The general conclusion that can be drawn from the analysis is that employing corpus linguistic tools to examine L2 pronunciation errors may enable one to create a thorough and reliable collection of commonly mispronounced words, which can constitute an effective and powerful tool in pronunciation teaching and learning. At the same time, careful examination of the corpus-based list and the process of its creation reveal that, just as in the case of compiling a list of common mispronunciations using "traditional" methods, creating a corpus-based index of pronunciation errors entails certain problems that need to be addressed when attempting to produce such a list.

1 Introduction

Sobkowiak's Words Commonly Mispronounced, provided in one of the appendixes of his *English pronunciation for Poles* (Sobkowiak, 2001), is a notable collection of over six hundred English words that are frequently mispronounced by Polish learners.

M. Zając (✉)
University of Łódź, Łódź, Poland
e-mail: zajac1234@gmail.com

E. Waniek-Klimczak and M. Pawlak (eds.), *Teaching and Researching the Pronunciation of English*, Second Language Learning and Teaching,
DOI 10.1007/978-3-319-11092-9_9

As explained by the author, the items on the list have been "(…) collected from experience as well as other books on English phonetics" (Sobkowiak, 2001, p. 350). Words Commonly Mispronounced (Sobkowiak, 2001) lists the troublesome English words together with their correct realisation (in this case, the standard British English pronunciation) and the most common erroneous realisation by Polish learners. The words are ordered by their frequency of occurrence, which seems to be based on frequency counts performed over a dictionary of English and over a stretch of running English text.[1] It is assumed that the words which can be found at the very top of the list will be problematic mostly for beginners, whereas the words placed closer to the end of the index may prove difficult to pronounce for advanced learners as well.

As observed by Szpyra-Kozłowska and Stasiak (2010, p. 3), "(…) although the list is placed in the appendix and thus is marginalized in the book, it belongs, as is often admitted, to the most frequently used parts of it." This is hardly surprising since an index of words that are commonly mispronounced by learners of English seems like a neat and simple way of substantially improving one's pronunciation in a relatively short amount of time and with relatively little effort. In English pronunciation teaching, such a list can supplement the practice of L2 segmental and prosodic features and can be employed to set teaching priorities, thus improving the effectiveness of instruction. One could also argue that a list of words that are commonly mispronounced could be easily integrated into English lessons at schools and be used by teachers who are less familiar or feel less comfortable with the English sound system. Finally, it can serve as an outside-of-school resource material that, arguably, could be utilized even by beginner learners of English with little phonetics and phonology training.

Even so, the list of Words Commonly Mispronounced (Sobkowiak, 2001) does appear to have a few weaknesses. First of all, it contains a number of rare and/or specialised pieces of vocabulary such as *plaid, vineyard, resound, nestle, bather, duplication*, etc. Since these items seem to occur in L2 speech relatively infrequently and, consequently, do not need to be given high priority in pronunciation teaching, they could perhaps be replaced with pronunciation errors that are more frequent in learner language.[2] Secondly, and more importantly, how is one to know which mispronunciations included in Words Commonly Mispronounced (Sobkowiak, 2001) or any other similar list are indeed the most common or frequent ones in learner speech? In fact, some seemingly notorious pronunciation errors could turn out to be isolated incidents that, for one reason or another, happened to catch phoneticians' attention. Rare pieces of vocabulary can be singled out as words commonly mispronounced simply because they are seldom used and

[1] However, the source of the frequency ordering in Words Commonly Mispronounced (Sobkowiak, 2001) is not clearly specified.

[2] Unless, of course, we take the phrase 'common mispronunciation' to mean that a given word is usually mispronounced whenever it is used, and not that it is frequently mispronounced in general. Still, a rare word that is mispronounced whenever it is used does not necessarily constitute a teaching or learning priority.

mispronouncing them is particularly striking and memorable. It is also possible that, for similar reasons, some common pronunciation errors go largely unnoticed.

Overall, although a list such as Words Commonly Mispronounced (Sobkowiak, 2001) is certainly an ingenious and useful tool for teaching and learning English pronunciation, an index that is based solely on a given phonetician or phoneticians' expertise does have certain limitations. These limitations, it would seem, could be overcome by adopting a corpus linguistic approach. Working on a representative corpus of learner speech should enable one to carry out a thorough analysis of pronunciation errors and, once these errors are identified, make it possible to quantify them easily. Thus, one should be able to compile an objective and reliable list of mispronunciations that are found to be the most frequent ones in learner language. An attempt to produce such a corpus-based index was made by Zając and Pęzik (2012), who used data from the spoken component of the Polish Learner English Corpus PLEC (http://ia.uni.lodz.pl/plec/). The aim of this paper is to describe the process of compiling a corpus-based index of frequently mispronounced words, discuss the most frequent pronunciation errors in the PLEC corpus, compare the corpus-based index with Words Commonly Mispronounced (Sobkowiak, 2001), and, finally, discuss the problems related to creating a corpus-based list of mispronunciations.

2 Compiling a Corpus-Based List of Mispronunciations

The following subsections discuss data collection for the PLEC corpus and the type of participants that were recorded, the process of defining a pronunciation error (which is central to the compilation of a corpus-based index of mispronunciations) and the format that was used to annotate the pronunciation errors in the corpus. The final subsection describes the results of the study by Zając and Pęzik (2012), i.e. the obtained list of the 50 most frequent mispronunciations in the PLEC corpus.

2.1 The Corpus

The Polish Learner English Corpus PLEC (http://ia.uni.lodz.pl/plec) comprises time-aligned interviews of Polish learners of English (200,000 words). The recordings were transcribed orthographically, time aligned and annotated for pronunciation errors. The participants were mostly advanced and intermediate learners of English (ranging from B1 to C2 proficiency levels); some speakers with elementary knowledge of English were also recorded (A1 and A2 proficiency levels). The group of participants consisted of students of English Studies recruited from the University of Łódź, secondary school students, junior high school students and adult learners. The subjects were interviewed about their hobbies and/or instructed to describe pictures and answer picture-related questions. The interviews were conducted alternately by a fifth-year student of English studies and three academic

teachers of English. The interviewers' speech was included in the analysis of pronunciation errors. The number of participants (together with the interviewers) totalled 130 speakers.

2.2 Defining a Pronunciation Error

A crucial stage of compiling a corpus-based list of frequent mispronunciations is coming up with a working definition of a pronunciation error that will be used to determine whether a particular realisation of a word should be treated as erroneous and included in the analysis. Zając and Pęzik (2012) decided to focus on pronunciation errors that involve segment substitutions and/or incorrect stress placement which are *not* caused by regular features of a Polish accent. Table 1 provides examples of the types of errors that were included/not included in the analysis.

The idea behind concentrating on mispronunciations that are *not* caused by regular features of a Polish accent was that, presumably, the resulting list would not require a detailed knowledge of the English sound system in order to understand the errors. This way, the list could be utilized not only by students of English studies or pronunciation enthusiast but also by teachers and learners who are less familiar with phonetics and phonology and/or learners of English for whom pronunciation is not a top priority.

The mispronunciations on the corpus-based list compiled by Zając and Pęzik (2012) are different from those listed in Words Commonly Mispronounced (Sobkowiak, 2001) in that the latter list includes mispronunciations that could be treated as resulting from regular features of a Polish accent (e.g., *which* pronounced with

Table 1 Types of pronunciation errors included and not included in the analysis in Zając and Pęzik's (2012) study

Excluded from the analysis	Included in the analysis
Polish-accented realisation of a vowel category (e.g., replacing the KIT vowel with Polish /i/, replacing the TRAP vowel with Polish /a/, replacing the THOUGHT vowel with Polish /o/, etc.)	Wrong stress placement (e.g., placing stress on the second syllable in *area,* placing stress on the first syllable in *event*, etc.)
Polish-accented realisation of a consonant category (e.g., realising dental fricatives as Polish /t d/, realising /h/as Polish /x/, realising dark /l/as clear /l/, etc.)	Replacing one vowel category with another (e.g., using the STRUT vowel in *butcher*, using the GOAT diphthong in *broad*, etc.)
Using full vowels in unstressed syllables (e.g., pronouncing the second syllable of *doctor* with a full vowel, etc.)	Replacing one consonant category with another (e.g., replacing /s/with /z/in *basic*, replacing /ʃ/with /ʒ/in *Croatia*, etc.)
Failure to maintain the voiced-voiceless contrast in English final obstruents (e.g., realising *eggs* as *ex*, *bag* as *back*, etc.)	Adding (or omitting) a sound (e.g., realising *lamb* with/b/, realising *debt* with/b/, etc.)

Polish /i/, *water* realised with Polish /o/, *suppose* pronounced with a full vowel in the unstressed syllable). Another difference is that a given realisation of a word was treated as erroneous if, as opposed to Words Commonly Mispronounced (Sobkowiak, 2001), it deviated from the pronunciation of said word in Standard Southern British English (SSBE) *and* General American (GenAm). In order to determine whether a given realisation can be found in either of these accents, Longman Pronunciation Dictionary (Wells, 2008) was consulted.

2.3 Annotating the Mispronunciations

The annotations of pronunciation errors were created manually by the author of this paper using ELAN Linguistic Annotator (http://www.lat-mpi.eu/tools/elan/; Sloetjes & Wittenburg, 2008). The time-aligned mispronunciations were marked using the following format: erroneous realization (IPA)|orthographic form|correct pronunciation (IPA), e.g. dɒnt|don't|dəʊnt. The mispronunciations were transcribed with the IPA symbols used for SSBE and not the symbols used for Polish (or GenAm), as it was assumed that it would facilitate the process of annotation. Also, when a given erroneous realization involved mispronunciations that were judged to have been caused by regular features of Polish accent, the mispronounced segments were transcribed as if they were realized native-like. For example, *Disney* pronounced as [ˈdʲisnej] would be transcribed as /ˈdɪsneɪ/, *foreign* realized as [fɔˈrejn] would be transcribed as /fəˈreɪn/. Similarly as in the case of SSBE phonetic symbols, this method of annotation was selected in an attempt to simplify the process.

2.4 The List of Frequent Mispronunciations

The complete list of fifty most frequently mispronounced words in the PLEC corpus can be found in the Appendix. The errors are arranged according to the total number of occurrences of a given mispronunciation in the corpus and the number of times a given word was mispronounced by different speakers. The results indicate that one of the most frequent types of mispronunciation in the PLEC corpus was replacing the GOAT diphthong with an open or mid back rounded vowel, as in *don't, old, also, Polish, only, Poland, older, told, photos, whole, most* and *moment*. Another common error was using an open or mid back rounded vowel in place of /ʌ/in *some, love, other, something, front, London, colours* and *company.* Similarly, an /o/-like vowel was often used to replace the NURSE vowel in *work, world* and *words*. Many items on the list of the 50 most frequent mispronunciations are words that have been misstressed by the participants, i.e. *kilometres, computer, interested, interesting, recommend, develop, exam(s), guitar, event(s), foreign.* Some other frequent errors include mispronouncing the digraphs < ei > and < ey > as in *their, volleyball* and *foreign*. The words *Warsaw* and *abroad* are also on the list,

illustrating that the participants often replaced the THOUGHT vowel with the GOAT diphthong. Other frequent errors in the PLEC corpus are the mispronunciations of the following words: *aren't, hobby, chemistry, biology, Czech, singing, education, fantasy, half* and *languages*. Overall, the list comprises many function words (e.g., *don't, their, some, also*) as well as words that are relatively infrequent in native-English corpora such as the British National Corpus and the Corpus of Contemporary American English (The British National Corpus, 2007; Davies, 2008), e.g. *volleyball, Warsaw, Czech* or *Polish.*

3 Discussion

The goal of the study by Zając and Pęzik (2012) was to provide a list of frequent mispronunciations in the spoken component of the Polish Learner English Corpus PLEC (http://ia.uni.lodz.pl/plec/). The following subsections discuss the results of the study. First, there is a more general discussion of the mispronunciations that were found to be the most common in the PLEC corpus; the possible sources of some of these errors and the possible reasons why certain types of errors appear on the list are described. Next, the list obtained by Zając and Pęzik (2012) is compared with Sobkowiak's (2001) Words Commonly Mispronounced.

3.1 General Discussion

The majority of the pronunciation errors on the list of 50 most frequently mispronounced words in the PLEC corpus seem to stem from inappropriate interferences from spelling, a trend that is especially visible in the case of the letter < o>, mispronounced as open to mid back rounded vowel, which is an approximation of this letter's realisation in Polish. Erroneous realisations of words such as *hobby, chemistry* and *Czech* seem to be strongly affected by Polish spelling conventions also. A similar phenomenon was observed, for instance, by Piske et al. (2002), who examined the realisation of English vowels by native Italian speakers and discovered that some participants' realisation of certain phones was affected by L1-inspired spelling conventions.

Some frequent mispronunciations in the PLEC corpus, on the other hand, appear to result from an overgeneralization of *English* spelling conventions, e.g. many participants realised *abroad* with the GOAT diphthong, probably due to the fact that the digraph < oa > is often realised as /əʊ/in English (*road, coast, coat, moan, goat, throat, load*, etc.). Other frequent errors on the corpus-based list (e.g., *volleyball, aren't, foreign, their*) appear to be linked to the words' spelling, but do not lend themselves to easy categorization. Regardless of the exact source of the spelling interference, the results of the study by Zając and Pęzik (2012) lend support to a statement made by Wells (2005, p. 104) that "[m]any oddities of the

NNS pronunciation of English are due to inappropriate interference from the spelling."

Another finding in the study by Zając and Pęzik (2012) is that many of the most frequent pronunciation errors in the PLEC corpus involve incorrect stress placement. Some examples of such mispronunciations are the words *computer* and *guitar* realised with primary stress on the first syllable. This tendency may be brought about by an overgeneralization of syntactic category based rules. Waniek-Klimczak (2002) and Archibald (1997) found that Polish learners of English tended to use word-initial stress in nouns, presumably because primary stress in English nouns frequently falls on the first syllable, and Polish learners extended this rule also to those lexical items to which it does not apply. Another source of stress-related errors may be L1 transfer. Stress in Polish is fixed on the penultimate syllable and this may be the reason why the participants realised words such as *ˈkilometres, ˈinterested, ˈinteresting, recoˈmmend, eˈvent(s)* and *eˈxam(s)* as *kiloˈmetres, inteˈrested, inteˈresting, reˈcommend, ˈevents* and *ˈexam(s).* The effect of L1 transfer on stress placement by Polish learners of English was also observed by, among others, Barańska (2011) and Matysiak (2012).

It was also found that many of the most frequent mispronunciations in the corpus-based list are function words (*don't, their, some, also, aren't*). This observation seems hardly surprising given the fact that function words occur frequently in speech in general. Nonetheless, the finding draws attention to the fact that function words should perhaps be given higher priority[3] in English pronunciation teaching. Since they appear in speech so often, they may lead to more breakdowns in communication and cause more irritation on the part of the listener than seemingly more serious errors that crop up in the learners' speech less frequently.

Finally, it seems worth mentioning that some of the items that are high on the list of the most frequent mispronunciations in the PLEC corpus seem to be relatively rare (e.g., *volleyball, Warsaw, Czech, Polish*). Obviously, words such as *Poland, Polish* and *Warsaw,* although not necessarily very common in the English language in general, are definitely frequent in the speech of Polish learners of English. As regards words such as *volleyball* and *Czech* as well as *hobby, kilometres, chemistry* or *biology* (which are also close to the very top of the corpus-based index of pronunciation errors), they seem to appear on the list partially due to the fact that many of the participants were secondary school students interviewed about their hobbies and interests. Many subjects were also asked about their hometowns and trips abroad, which seems to explain why the list includes the words *Czech* and *kilometres*.

[3] Admittedly, when learners are mastering weak forms in English pronunciation classes, function words are the focus of much attention. At the same time, native-like pronunciation of function words which do not typically have weak forms (such as *don't, aren't, their*) may receive less attention.

3.2 Words Commonly Mispronounced Versus a Corpus-Based List

The items on the list of 50 most frequently mispronounced words in the PLEC corpus (Zając & Pęzik, 2012) are, for the most part, considerably different from the first 50 items in Words Commonly Mispronounced (Sobkowiak, 2001). The two indexes overlap only in a few instances: *their* (interestingly, it's the third item on both lists), *other, only, old* (*older* in the corpus-based list), *work, world, don't, money, half* and *front.*[4] The fact that these words have been spotted both by Sobkowiak (2001) and in the study by Zając and Pęzik (2012) implies that their pronunciation is especially difficult to master for Polish learners of English and/or that these mispronunciation are especially annoying for the listeners.

Nevertheless, the number of mispronunciations that are exclusive to only one of the lists is far greater than the number of errors which appear on both. For instance, the corpus-based list (Zając & Pęzik, 2012) includes words such as *some*, *love, aren't computer, interesting, colours, exam*(s), *develop* or *foreign*, which do not appear among the first 50 items in Words Commonly Mispronounced[5] (Sobkowiak, 2001). The first 50 items in Words Commonly Mispronounced (Sobkowiak, 2001), on the other hand, comprise items such *as said, saw, answer, heard, south, area* or *special*, which are absent from the list of 50 most frequently mispronounced words in the PLEC corpus (Zając & Pęzik, 2012).

Naturally, one needs to bear in mind that the types of errors included in the two lists are slightly different, i.e. some of the mispronunciations mentioned by Sobkowiak (2001) would be regarded as instances of regular features of a Polish accent and would consequently be excluded from the analysis in the study by Zając and Pęzik (2012). Moreover, the ordering of the words by frequency is not the same in the two lists. On the whole, however, the observations made here seem to validate the claim that adopting a corpus linguistic approach can be particularly advantageous in the examination of frequent pronunciation errors. At the same time, given the differences in data collection and the exclusion of Polish-accented errors from the study by Zając and Pęzik (2012), the similarities that can be found between the two lists do seem quite striking and intriguing,

4 Problems

Although a list of frequent mispronunciations produced with the use of corpus linguistic tools does have a number of advantages, the study by Zając and Pęzik (2012) revealed that there are also a number of problems related to the creation of

[4] Notice that, in the whole subsection, the author is comparing the first 50 words on the corpus-based index and the first 50 words on Words Commonly Mispronounced, not the complete lists.

[5] However, Words Commonly Mispronounced as a whole do contain *foreign, aren't* and *development*.

such a list and the creation of a thorough and reliable index of pronunciation errors in general. These issues include corpus representativeness, annotation format, maintaining objectivity when analysing the data, and, finally, the definition and classification of pronunciation errors.

4.1 Representativeness

As referred to in the first section of this paper, working on a representative corpus of learner speech should enable one to carry out a comprehensive analysis of pronunciation errors. However, collecting a representative database of learner speech is, in fact, not a simple task. The spoken component of the PLEC corpus seems fairly sizeable, but can it be considered representative of the spoken English of Polish learners? As explained by Biber (1993, p. 243),

> Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language.

The spoken component of the PLEC corpus includes only one type of linguistic text, i.e. spoken interactions between an interviewer and one or two learners. The range of linguistic distributions is also relatively limited; the conversation topics are mostly the same in all of the interviews, which resulted in the appearance of words such as *volleyball* or *hobby* at the very top of the list of the most frequent mispronunciations. All in all, although the spoken component of the PLEC corpus can definitely be of much use in the study of learner speech, working on a more diversified database should yield results that are representative of more than one type of linguistic context.

4.2 Annotation Format

In the study by Zając and Pęzik (2012), the authors decided that the mispronunciations should be transcribed with the IPA symbols used for SSBE and that if a realisation that was considered erroneous involved the use of some regular feature of a Polish accent, the segment(s) containing the Polish feature would be transcribed as if it was realized native-like (see Sect. 2.3). This annotation format was selected, because it was assumed that it would facilitate the process of transcription. It transpired later that it was, in fact, complicated and confusing. As a result, the transcriptions of the mispronunciations are sometimes inconsistent with each other and often do not reflect the actual realisations by the learners. For instance, the fact that the erroneous realisation of the word *Polish* is /ˈpɒlɪʃ/ (see Table A.1 in the

Appendix) does not mean that every subject pronounced the second vowel native-like, i.e. using the KIT vowel. Many participants realised it as a Polish /i/, but since the lack of a FLEECE-KIT contrast was considered a regular feature of a Polish accent (and was excluded from the analysis), it was not marked in the transcription. Also, if one is to follow the annotation guideline that elements in which a regular feature of a Polish accent was used should be transcribed as if they were pronounced native-like, one should transcribe [ˈdivelop], a common mispronunciation of the word *develop*, as /ˈdɪvələp/ (full vowels transcribed as schwas in accordance with the rule that vocalic elements are usually reduced in unstressed syllables in English). However, in this case, /ˈdɪvələp/is a far cry from the actual pronunciation of the word and does seem somewhat artificial. For this reason, realisations such as [ˈdivelop] were often transcribed as /ˈdɪvelɒp/. This type of transcription seems more natural, but is not in line with one of the annotation guidelines, resulting in transcription inconsistencies. The conclusion that can be drawn from these observations is that, first of all, the mispronunciation annotation format is a key element in a corpus-based examination of pronunciation errors, and, secondly, it is vitally important that the selected format is relatively simple to follow and, at the same time, reflects the actual realisations of the learners.

4.3 Objectivity

As mentioned earlier in this paper, one of the problems with lists of frequent mispronunciations that are collected from experience is the fact that seemingly prevalent errors can in fact be isolated incidents that happened to catch one's attention. Nonetheless, a similar kind of problem can arise when one is compiling a corpus-based list of mispronunciations. It was only one person that indentified the pronunciation errors in Zając and Pęzik's (2012) study, which is clearly insufficient to ensure completely objective judgements. In such a case, one cannot be absolutely certain whether the rater is not focusing on particular types of errors and overlooking others. Another factor that needs to be taken into consideration is mental fatigue (inevitable when annotating for several hours, as is usually the case), which can substantially reduce one's ability to single out pronunciation errors. In conclusion, it needs to be stressed that in order to produce a truly thorough and reliable index of the most frequent mispronunciations, several raters should be involved in the annotation process. Fortunately, with a database such as the spoken component of the PLEC corpus, a number of different people can easily listen to the same recordings. This way, the raters can check up on one another to increase objectivity and share the workload to avoid mental fatigue.

4.4 Definition and Classification of a Pronunciation Error

Some of the most interesting issues that arose during the examination of pronunciation errors in the study by Zając and Pęzik (2012) are the questions of how to definite a pronunciation error and how to classify a given mispronunciation. As referred to previously, Zając and Pęzik (2012) resolved not to concentrate on mispronunciations that are caused by regular features of a Polish accent and one of the reasons behind it was that the resulting list would likely appeal to learners who wish to improve their pronunciation but do not consider it a top priority. It was assumed, perhaps somewhat naively, that the very top of this list would comprise several serious pronunciation errors that can impede successful communication (thus rendering the list suitable and interesting for different types of learners, not limiting it to English students who are particularly interested in pronunciation). What the very top of the list actually contains are mostly mispronunciations that could possibly cause some irritation on the part of pronunciation teachers or native listeners. It is hard to imagine, however, that they would cause major breakdowns in communication. Moreover, realising words such as *love* with some sort of an/o/ seems perfectly acceptable in many regional accents of English. These observations suggest that, contrary to the authors' initial assumptions, it might prove more rewarding to take typically Polish pronunciation features into account. Indeed, the inability to differentiate between FLEECE and KIT or, for instance, a failure to maintain the voiced-voiceless contrast in English final obstruents could potentially prevent successful communication. Yet if one is to equate incorrect pronunciation with producing sounds that deviate from the native language norm, where should one stop? To what extent should the non-native realisation deviate from the native norm to be considered an error and what *is* the native language norm? The former question is especially important in the case of vowels, sounds that form a continuum with no distinct boundaries between one category and another. A given vowel category can cover a range of qualities, which, in some cases, could render it impossible to determine whether a given realisation is 'correct' or not. As regards the latter question, it can be difficult to decide whether a certain realisation is erroneous even when the native language norm is simply taken to mean standard pronunciation. For example, since the TRAP vowel can have quite distinct realisations in General American and Standard Southern British English (Lindsey, 2012), what sort of realisations of this vowel should be regarded as deviations from the norm? One might also wish to take regional accents into consideration, which could complicate matters even further.

Another problem is that erroneous realisations can often prove difficult to classify regardless of the criteria that are used to define a pronunciation error. For example, mispronunciations of words such as *told* or *old* seem to stem from a simple substitution of the LOT vowel for the GOAT vowel. Yet, words like *told* or *old* do not necessarily have to be produced with the sequence [əʊɬ]. A native speaker can also pronounce them with [ɒʊɬ] (Wells, 2008), which, in turn, is perceptually close to [ɒɬ] (the [ʊ]-type resonance of the following velarised approximant should render the

two realisations very similar). Thus, *told* and *old* produced with the LOT vowel seem like perfectly legitimate pronunciations. This signifies that the problem with native-like realisation of words such as *told* and *old* does not lie in the fact that learners are replacing GOAT with LOT, but rather in the fact that they are using the wrong allophone of /l/. Another mispronunciation that can prove fairly difficult to classify is realising *some* with an/o/-like vowel. On the one hand, it could be considered as a simple case of using the LOT vowel in place of the STRUT vowel. On the other hand, since the word *some* is usually pronounced with an unstressed, reduced vowel, one could treat the mispronunciation as an instance of a lack of vowel reduction. Erroneous realisations of words such as *certain*, *comfortable* or *determine* are similar in this respect. It is not clear whether producing a diphthong rather than a reduced vowel in the final syllable should be viewed as a problem with vowel reduction or as a result of inappropriate inference from the spelling.

Finally, it should also be mentioned that the definition of a pronunciation error employed in the study by Zając and Pęzik (2012) is not completely watertight. As mentioned earlier, the definition states that only the errors that are not caused by regular features of a Polish accent should be included in the list of mispronunciations. However, some of the most frequent pronunciation errors in the PLEC corpus, e.g. realising the < o > letter as /o/in words such as *don't, some, work, love, Polish*, could hypothetically be treated as a regular feature of a Polish accent (after all, this type of mispronunciation was very common among the subjects). The term 'regular features of a Polish accent' is clearly too broad and perhaps a better solution would be to prepare a more precise list of features that one wishes to exclude from the analysis.

5 Conclusions

The results of the study by Zając and Pęzik (2012) indicate that employing corpus linguistic tools to examine L2 pronunciation errors makes it possible to create new and improved "lists of words commonly mispronounced." A corpus-based list of frequent pronunciation errors can constitute an effective and powerful tool in pronunciation teaching and learning, especially since the researcher no longer needs to rely on anecdotal evidence in order to determine which mispronunciations are particularly common in learner speech. At the same time, one should bear in mind that compiling a corpus-based index of pronunciation errors is not without its difficulties. Before we set out to produce such a list, issues such as the definition of a pronunciation error or the representativeness of the corpus should be carefully considered.

Acknowledgments The study by Zając & Pęzik (2012) is part of a research project funded in the years 2010-2012 by a grant from the Polish Ministry of Science and Higher Education (N N104 205039).

Appendix

See Table A.1.

Table A.1 50 most frequently mispronounced words in the PLEC corpus together with sample erroneous realisations

	Word	Total no.	No. of speakers	Realisation
1	Don't	523	104	dɒnt
2	Old	113	35	ɒld
3	Their	94	26	ðeɪr, ˈðeɪʌr, ðeɪ
4	Some	138	23	sɒm
5	Work	73	22	wɔːrk, wɔːk
6	Also	80	21	ˈcːlsɒ, ˈɔːlzəʊ, ˈɔːlzɒ
7	Polish	78	21	ˈpɒlɪʃ
8	Poland	66	18	ˈpɒlənd
9	Love	68	16	lɒv
10	Something	61	16	ˈsɒmθɪŋ
11	Volleyball	50	16	ˈvɒleɪbɔːl, ˈwɒlibɔːl, ˈwɒleɪbɑːl
12	Only	55	14	ˈɒnli
13	Other	55	14	ˈɒðər, ˈɒðə
14	Front	53	13	frɒnt
15	Older	32	12	ˈɒldər
16	World	49	12	wɔːrd, wɔːrld
17	Hobby	32	11	ˈhɒbbi
18	Kilometres	32	11	kɪləˈmiːtərz, kɪləˈmiːtəz, kɪləˈmetərz
19	Aren't	32	10	ˈɑːrənt
20	Computer	49	10	ˈkɒmpjuːtə, ˈkɒmpjuːtər
21	Interested	33	10	ɪnˈtrestɪd, ɪntəˈrestɪd
22	London	25	9	ˈlɒndən
23	Warsaw	30	9	ˈwɔːrsəʊ, ˈwɜːrsəʊ
24	Chemistry	38	8	ˈhemɪstri, ˈʧemɪstri
25	Interesting	23	8	ɪnˈtrestɪŋ, ɪntəˈrestɪŋ
26	Abroad	16	7	əˈbrəʊd
27	Biology	18	7	ˈbjɒlədʒi, baɪɒˈlɒdʒi
28	Czech	18	7	ʧeh
29	Recommend	29	7	rəˈkɒmend, ˈrekəmend

(continued)

Table A.1 (continued)

	Word	Total no.	No. of speakers	Realisation
30	Most	14	6	mɒst
31	Singing	40	6	ˈsɪŋɪŋg, ˈsɪŋgɪŋg
32	Told	14	6	tɒld
33	Colours	16	5	ˈkɒlərz
34	Company	22	5	ˈkɒmpəni, kɒmˈpæni
35	Develop	17	5	ˈdɪvelɒp
36	Education	17	5	eduˈkeɪʃn
37	Exam	20	5	ˈegzæm
38	Exams	20	5	ˈegzæmz
39	Fantasy	12	5	fənˈtæzi
40	Photos	25	5	ˈfəʊtɒz, ˈfɒtɒz
41	Whole	22	5	hɒl
42	Words	13	5	wɔːrdz
43	Guitar	14	5	ˈgɪtɑː, ˈgɪtɑːr
44	Event	12	4	ˈiːvənt
45	Events	15	4	ˈiːvənts
46	Foreign	11	4	fəˈreɪn, ˈfɔːreɪn
47	Half	9	4	hɑːlf
48	Languages	13	4	ˈlæŋgwɪʃɪz
49	Moment	11	4	ˈmɒmənt
50	Money	10	4	ˈmɒneɪ,ˈmʌneɪ

References

Archibald, J. (1997). The acquisition of English stress by speakers of nonaccentual languages: lexical storage versus computation of stress. *Linguistics, 35,* 167-181.

Barańska, A. (2011). The examination of the vowel length in the pronunciation of the English adjectives with -able, -ate, -ative suffixes. *Gavagai Journal, 1*, 3-16.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8,* 243-257.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca/. Accessed 28 April 2014.

Lindsey, G. (2012). The British English vowel system. Speech Talk. Thoughts on English, speech & language. http://englishspeechservices.com/blog/british-vowels. Accessed 28 April 2014.

Matysiak, A. (2012). Is English word stress beyond the scope of Polish advanced learners of English? – different strategies of indicating the stress position within multi-syllable words. *Gavagai Journal, 2,* 22-33.

Piske, T., Flege, J. E., MacKay, I., & Meador, D. (2002). The Production of English Vowels by Early and Late Italian-English Bilinguals. *Phonetica, 59,* 49-71.

Sobkowiak, W. (2001). *English phonetics for Poles*. (2nd ed.) Poznań: Wydawnictwo Poznańskie.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008).

Szpyra-Kozłowska, J., & Stasiak, S. (2010). From focus on sounds to focus on words in English pronunciation instruction. *Research in Language, 8,* 1-12,

The British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/. Accessed 28 April 2014.

Waniek-Klimczak, E. (2002). How to predict the unpredictable – English word stress from a Polish perspective. In E. Waniek-Klimczak, & P. J. Melia (Eds.), *Accents and speech in teaching English phonetics and phonology* (pp. 221-242). Berlin: Peter Lang.

Wells, J. C. (2005). Goals in teaching English pronunciation. In K. Dziubalska-Kołaczyk, & J. Przedlacka (Eds.), *English pronunciation models: A changing scene* (pp. 101-112). Bern: Peter Lang.

Wells, J. C. (2008). *Longman Pronunciation Dictionary*. Harlow: Pearson Education Limited.

Zając, M., & Pęzik, P. (2012). Developing a corpus-based index of commonly mispronounced words. Paper presented at *Accents 2012 - VIth International Conference on Native and Non-native Accents of English*, Łódź, Poland, 6-8 December, 2012.