This is the accepted manuscript of the article, which has been published in Tyrkkö J., Nurmi A. & Tuominen J. (eds) Challenging the Myth of Monolingual Corpora. Leiden: Brill. Language and Computers no. 80. ISBN: 978-90-04-27668-0. ISSN: 0921-5034

Semi-automatic discovery of multilingual elements in English historical corpora: Methods and challenges

Jukka Tyrkkö, Arja Nurmi and Jukka Tuominen

1 Introduction

One of the main obstacles standing in the way of large-scale studies of multilingual practices has been the difficulty of discovering secondary, i.e., foreign or minority language words by any other than manual means (see Pahta 2004; 2007; 2011; Nurmi and Pahta 2012). Although current computational methods for identifying the primary language of a monolingual text are robust (see, e.g., Alex, Dubey and Keller 2007), their accuracy diminishes significantly when the task is to identify short chunks of words and phrases in other languages within texts written predominantly in English (see King, Kübler and Hooper 2015). While established language-detection methods such as the use of foreign-language tags in POS-annotated corpora (see Grefenstette 1995), dictionary-based retrieval (see Alex 2005), and non-standard letter clusters (see Andersen 2012) can all be useful in word-level language identification, acceptable rates of both precision and recall can only be accomplished when the discovery logic is based on triangulation between multiple methods at the same time. Furthermore, when the objective is not only to identify passages that appear foreign based on form alone but also to distinguish between switches into another language and word sequences that have been assimilated into contemporary English usage in a particular register, the task becomes significantly more challenging.

In this paper we discuss the semi-automatic method of discovering foreign-language passages developed by the *Multilingual Practices in the History of Written English* project[1] for the purpose of corpus-based analysis of multilingual practices in English historical texts (see Nurmi et al. forthcoming; Tyrkkö and Nurmi forthcoming). Using the *Corpus of Late Modern English Texts 3.0*

[1] The project was based at the University of Tampere and funded by the Academy of Finland (2012–2016, project number 258434). The principal investigator was Päivi Pahta, senior scholars were Arja Nurmi, Laura Wright, Jukka Tyrkkö and Janne Skaffari, and junior scholars Jukka Tuominen, Anna Petäjäniemi and Veera Saarimäki.

(CLMET3.0) as a training corpus, we developed Multilingualiser 1.0, a proprietary tagger tool that uses multi-dictionary look-up, collocate analysis, and character n-grams to identify words that have a high probability of being foreign. Importantly, the mere use of foreign-looking words and phrases does not invariably mean that a foreign language is actually being used, and consequently the automatic discovery phase was followed by post-discovery evaluation, where experienced members of the project team verified the automatically assigned language tags, confirmed sequence lengths, and discarded items that were correctly identified but evaluated not to be evidence of multilingual practices in that specific register. The expert-evaluation phase will be explained in detail, including discussion of the taxonomy of foreign-language passages. The discussion will be illustrated with examples from the corpus data.

Section 2 defines what we mean by multilingual practices and discusses earlier corpus-linguistic studies of the phenomenon. In section 3, we describe the process of identifying multilingual practices in CLMET3.0, both in terms of automatic identification and human post-processing. Section 4 concludes the paper.

2 Multilingual practices and corpus linguistics

In our study, we define multilingual practices as "the use of more than one language in the course of a single communicative episode" (Heller 1988, 1). While Heller's definition refers specifically to code-switching, we deliberately avoid that term, partly because it has a strong association with the study of spoken language and partly because of the frequent limitation of code-switching as a term related to bilingualism through language acquisition rather than language learning. Our data consists of written language and the multilingual elements in it are typically evidence of languages the writers have learned at school, not acquired as members of a linguistic minority community. Having said that, we do refer to switched passages, switch points and switching in our paper, but

this is due to practicality, not to a theoretical alignment with any definition of the term *code-switching*.

Diachronic multilingualism is still a developing field of inquiry, where a particular focus of interest, especially of late, has been testing the hypothesis that multilingual practices are a pervasive if not ubiquitous feature of language use (see, e.g., Schendl and Wright 2011). As far as English is concerned, recent scholarship supports the argument that multilingualism has been a consistent feature throughout the history of the language and that multilingual practices have been particularly characteristic to specific domains such as medicine and science, religion, commerce and administration (see, e.g., Voigts 1989; Schendl 1996; Wright 1999; Pahta 2004; Davidson 2005; Pahta and Nurmi 2006; 2011).

One of the latest methodological developments in historical multilingualism studies is the application of corpus linguistic methods to corpora that were originally compiled for the study of historical varieties of English in an effort to discover the extent to which foreign words and phrases were used by native speakers of English writing to an audience of other native speakers. Pioneered by Nurmi and Pahta (2004; Pahta and Nurmi 2006; 2011), this approach holds the dual promise of allowing us to establish baseline frequencies for foreign-language elements in English texts and to discover correlations between multilingual practices and sociolinguistic and text typological variables.

The present study was produced under the aegis of the *Multilingual Practices in the History of Written English* project, and is part of our larger research project on the methodological challenges of identifying and quantifying multilingual passages in historical corpora. Addressing the topic both in general terms and with particular reference to our study of the 34-million-word CLMET3.0 (Nurmi et al. forthcoming), we introduce the *Multilingualiser* tool, discuss the methods

we have developed for tagging foreign-language passages in corpora and the key issues of quantifying multilingual practices.[2]

Our project studies multilingual practices in historical material from the two primary perspectives of frequency and usage. Currently, we are primarily focused on foreign-language passages in written and published English texts. The use of more than one language is a frequent and essentially expected phenomenon in spoken discursive contexts among bilingual speakers of minority languages within their own linguistic community. It is also found where the participants are not members of the same community, such as language contact and English-as-Lingua-Franca situations (see, e.g., Lange, this volume; Hynninen, Pietikäinen and Vetchinnikova, this volume). More relevantly for our study, several languages may also be used within discourse communities consisting mostly or even entirely of native speakers and writers of one language (see, e.g., Androutsopoulos 2011; Leppänen et al. 2011; Kytölä 2012; see also Kaunisto, this volume, discussing these questions through examining a particular text included in CLMET3.0). We posit that such practices are more deliberate and measured in written than spoken discourse, the former affording more time for processing and consideration. Therefore, the use of foreign words and phrases in primarily English text can be approached as a purposeful discursive choice. Notably, this goes counter to what ethnographic studies show us, namely that switching to another language is often perceived to be lazy language use, that speakers (including those who do so) disapprove of the practice, and that speakers are generally not aware of the extent to which they themselves make use of several languages (Gardner-Chloros 2011, 14–15). Therefore, any analysis of multilingual practices in written texts ought to take into account whether or not the author is likely to have considered the words or phrases they use as foreign, either to themselves or to their readers. After all, the markedness of multilingual practices depends considerably on the community in question (cf., e.g., Poplack 1988; Myers-Scotton 1993): while switching between several languages can be

[2] We use the term *foreign* because it suits our data, where most evidence of multilingual practices consists of the use of languages that are in that category. This is not to say that the Multilingualiser could not be applied to the identification of, for example, minority languages.

the communicative norm and attract hardly any attention in a highly heterogeneous community such as at a busy trading post or in a scholarly treatise, the use of a single foreign word may be highly noteworthy in an intensely monolingual community or in a book written for young children. Different languages will likewise elicit dissimilar reactions depending on the context: phrases in Latin may serve a genuine communicative purpose in academia or in the field of medicine, whereas the same phrases may mark the speaker as a terrible snob or a bore in a more leisurely setting. Because CLMET3.0 is based entirely on published texts, the intended readers of the texts can – and should – also be considered as customers. The sociolinguistic and text typological aspects of the data will be discussed in much greater depth in Nurmi et al. (forthcoming).

In practice, our analysis of multilingual practices begins with the identification and quantification of foreign words and phrases, and then proceeds to a manual pragmatic analysis where we verify the nature of the passages and identify the general type of each switched passage using a tripartite model of *conventionalised*, *pre-fabricated* and *free* switches. By conventionalised switches we refer to short, formulaic expressions which are familiar to writers and readers without any significant knowledge of the language in question (*au revoir*, *ad nauseam*); they can often be considered to occupy the grey area between switching and borrowing. Pre-fabricated passages consist of quotations, proverbs and maxims (*alea iacta est*): more complex elements, which nevertheless do not require the writer to produce original stretches of language. Finally, free switches are the type most closely approximating the linguistic practices of fluent multilinguals: elements in each of the languages included in an utterance are produced by the writer (for further details, see Nurmi et al. forthcoming and section 3.3 below).

Traditionally, linguistic inquiry has focused predominantly on one language at a time, with preference being shown to language produced by native speakers, and consequently many corpus

compilers, too, have opted to discard texts that contain copious passages in foreign languages.[3] This is particularly true of extract corpora – i.e., corpora comprising short extracts rather than full length texts – where it makes good sense to select passages that include only the language one is principally interested in studying. Many compilers do not explicitly address issues of multilingualism in corpus manuals or articles discussing the compilation process. Given that English corpora have generally been designed to be representative of English exclusively, any evidence of multilingualism discovered in such corpora may have been perceived by the corpus compilers to be acceptable as English. On the other hand, the design principles may have led the compilers to discard texts that contain unacceptably high quantities of non-English writing. Quantitative evidence of multilingualism drawn from conventional monolingual corpora may therefore be somewhat skewed, leading to less frequent multilingual practices than is perhaps found in the language at large.

In historical corpus linguistics, issues of multilingualism are at once more pronounced and more complicated. It is well-known that the further back we go, the less extant data is available and the more narrow the sociolinguistic strata from which the texts originate. The overall representativeness of historical corpora is therefore generally much weaker than that of present-day corpora, and more philological sensitivity is required when conclusions are drawn from them. In order to include as much data as possible, compilers will often choose to include texts that are not as monolingual as they might ideally want, but equally multilingualism may be recognised as an important feature to be represented, possibly with appropriate flagging of content; for example, the *Middle English Medical Texts* (MEMT) corpus includes multilingual verse as a separate category (see also Rütten, this volume, on multilingualism in the *Corpus of English Religious Prose* (COERP)).

[3] A notable shift in this practice can be observed in so-called megacorpora compiled through web and archive crawling. Typically, the inclusion of an undefined amount of non-English text in megacorpora is dismissed as "noise" and as an artefact of the compilation process.

When included in a corpus, foreign language content may be flagged in the metadata or explicitly annotated into the corpus itself. Both require some extra analytical work on the part of the compiler, the latter particularly so. In historical corpora, tagging of foreign content has been typically carried out manually (see, e.g., Nurmi and Pahta 2004); examples of corpora with annotation of foreign words include the *Helsinki Corpus*, the *Lampeter Corpus* and the *Parsed Corpus of Early English Correspondence*. However, when it comes to large-scale quantitative analysis of multilingual practices with multi-million-word datasets that yield sufficient evidence for inferential sociolinguistic and text typological analysis, the amount of work required will quickly become unmanageable. While it is possible to annotate corpora of any size manually, doing so would require such a massive investment of time and effort that few research projects can afford it. Given our objective of using the 34-million-word CLMET3.0 corpus, the only logical solution was to turn to computational methods for help.

3 Defining and identifying foreign language use in CLMET3.0

The identification of multilingual elements involves a variety of theoretical decisions at the best of times and the challenges are multiplied when the process is to be carried out automatically. The task can be broken down into four components. First, we must arrive at a working definition of multilingual practices that does not leave too many fuzzy areas. Secondly, we need to consider the problems of recognising and retrieving multilingual elements from the corpus. The third step is operationalising the computational discovery process, which in this context means the Multilingualiser software package. Finally, the post-processing of retrieved results must be carried out by non- or semi-automatic means, relying on human expertise.

3.1 Problems of recognising and retrieving multilingual elements

The first decision to be made at the very beginning of any scripting process is to define what, exactly, we want the software to find or, in our particular case, what we mean by multilingual practices. We may well prefer a flexible, resource-focused definition of multilingual practices when discussing the functions of switching based on data from smaller corpora, but when it comes to programming we have to be very specific about what it is we want to find. We may theoretically agree that cut-and-dried definitions are difficult and even counter-productive to provide, but corpus-linguistic and quantitative analyses can only be carried out if we can operationalise a method for distinguishing foreign content from the rest of the text. One approach is to rely on the intuition of a mature, competent speaker of the language in question – or, preferably, more than one such speaker. However, although expert opinions are valuable and we did rely on them for the post-discovery phase of the process, they cannot be the basis of a computational model intended to process large corpora. So how did we identify foreign words in the data?

A number of criteria have been proposed in literature. One commonly used criterion is phono-morpho-syntactic, so that there is no morphological, phonological or syntactic convergence into the system of the main language of the text (see, e.g., Poplack 1988). However, although this would mean in the strictest sense that any orthographically un-anglicised word form constitutes a switch, this is clearly not true. For example, while it may be argued that the use of a singular or plural form of a Latinate word in English may be considered a stylistically marked form, such as the singular *datum* for *data*, there are plenty of cases where such a practice does not signify status as a foreign word, e.g. *phenomenon*/*phenomena*, *corpus*/*corpora*. Even such a seemingly simple criterion cannot, therefore, be seen as unambiguously applicable.

Other useful criteria may be frequency of use or first date of use in the target language, though here we need to be careful to distinguish between registers and communities of practice: for example, a word or phrase commonly used among academics or within the small circles of a group of hobbyists may be almost entirely unknown to the general public.

A key challenge in the discovery of foreign content is the identification of relevant items from the text and the evaluation of whether or not each individual occurrence counts as a switch to another language. Although these are two separate issues, they cannot be treated entirely apart. In technical terms, we are dealing with a pattern recognition problem and thus the concepts of *precision* and *recall*. Do we want to retrieve every item that could possibly be foreign, only those items that are definitely foreign, or something in between? The first option would mean high recall and low precision, and require excessive amounts of manual processing post discovery, while the latter would mean the opposite and lead to a potentially high volume of missed items and consequently to systematic errors in the actual analysis.

In addition to the problem of deciding which words are foreign, there is the challenge of identifying those words from normal running text. Automatic language detection is a well-known and to a large extent well-met challenge in information retrieval and computational linguistics. Sorting documents into groups based on their dominant language is a machine-learning task that can be performed with very good accuracy using a variety of computational methods. However, things become much more challenging when we are dealing with single foreign words or short passages of words occurring within otherwise ostensibly monolingual running text (cf. Alex 2005) – to say nothing of the fact that those foreign words can be in any number of different languages and can, as evidence now shows us, occur in a variety of syntactic positions with little to no explicit flagging. Switching can occur intersententially and intrasententially at virtually every level of language from single morphemes within a lexical item to passages substantial enough to be considered self-contained discourses-within-discourse, such as an entire short story embedded within another text (see also Kaunisto, this volume, and Kohnen, this volume). Consider, for example, the following examples:

(1) Among the English who this year elected to take their repose and recreation at Trouville there was no more brilliant figure than Mrs. Luke Widdowson. This lady is well known in

the *monde* where one never *s'ennuie*; where smart people are gathered together, there is the charming widow sure to be seen. (Gissing, *The Odd Women*)

(2) […] the Grand Duke of Pfeifentopf says: "You have me with your writings much refreshed. I have the whole revenues of the Grand Duchy against one thousand *flaschen* of *lager bier gebetted*, and I have won him on your noble advice on Marvel." (*Punch* magazine, vol. 99)

In example (1), the French word *monde* is provided with an English definite article and the reflexive *s'ennuie* carries on directly from 'where one never', taking the English pronoun 'one' as its subject. The CLAWS5 part-of-speech tagger, for example, catches *monde* as a foreign (or unclear) word but tags *s'ennuie*, despite its decidedly un-English form, as a noun.

In example (2), the German noun *flaschen*, 'bottles', has the correct plural declension as it follows the English numeral 'thousand'. It is followed by the English preposition 'of', immediately followed by yet another potential switch to German with *lager bier gebetted*. In this instance at least *bier* but perhaps also *lager* must be considered German, but what about *gebetted*? Although the word looks German with the past participle marker *ge-*, and its position at the end of the clause follows German syntax, it has the English ending *-ed* (instead of the German *-et*), and the verb itself would be wrong for the context: *betten* does not mean 'betting' in the sense of 'gambling'.[4] The word *gebetted* is a linguistic joke in faux-German or a comic representation of faulty German-influenced English. It clearly demonstrates a certain linguistic prowess, but at the same time is not, in fact, a free switch to any other language.

To make the detection of foreign words even more difficult, the insertions often come from languages that are morphologically quite similar to English, and in many cases at least some of the words in longer passages are perfect *visual diamorphs* of existing English words – that is, they look exactly like English and thus cannot be identified by simple dictionary look-up procedures (see Wright 2011; Stam and de Schepper forthcoming). Examples of particularly frequent such items

[4] The German cognate would be *wetten*, while *betten* translates as 'to lie/set down, embed'.

include *a* and *an*, as in, for example, *a priori* or *an sich*. However, much more challenging are cases such as examples (3) through (5):

(3) […] these tombs were not all equally well cared for (*Post mortem nescio!*) and it had been one of the pieties of Aurelius to frame a severe law to prevent the defacing of such monuments. (Pater, *Marius the Epicurean*)

(4) […] no farther will I go than the duties of my *post oblige* me, and that honour, which to forfeit, would render me unworthy of your care. (Haywood, *The Fortunate Foundlings*)

(5) Still, in his present humour, Dickie's sense of *noblesse oblige* was strong. (Kingsley, *The History of Sir Richard Calmady*)

The word *post* appears in examples (3) and (4), and the word *oblige* in examples (4) and (5). Notably, while in example (4) both words are English, *post* in example (3) and *oblige* in example (5) are Latin and French, respectively. Determining this is reasonably simple for a competent human reader, but if an automatic tool is expected to accomplish the same task, it must have the ability to analyse the collocates as well as the immediate multi-word units, for example, to catch *noblesse oblige* as a two-word unit of French rather than a French noun followed by an English verb or to understand that *post*, *oblige* and *me* in example (4) are not foreign words. Accordingly, the tagger needs to take into consideration that all three are also English words, that an intrasentential string with one Latin word followed by two French words is highly unlikely, and that two of the three – *post* and *me* – are function words in two different languages, making the combination ever more impossible.

The task of annotating foreign content is thus challenging in several different ways. The tagger will need to be sensitive to different languages, it will need to be aware of the fact that words sometimes occur in multi-word units, that function words behave differently than content words, that some words in different languages may look the same, and that foreign words of the same language tend to co-occur.

3.2 Operationalising the computational discovery process: Multilingualiser

Multilingualiser is a corpus tool designed to speed up the task of automatically annotating possibly foreign words and phrases in corpora. Written by Jukka Tyrkkö in LiveCode, a high-level programming language available open-source for non-profit use,[5] Multilingualiser bundles together a series of different foreign-language detection methods, a tag editor, and a chunking tool. Multilingualiser comes with an easy-to-use graphic user interface, requires no installation and can output data in several widely used formats.[6]

Because hardly any frequency data and few theoretical models were available on multilingual practices in Late Modern English, the development of the software proceeded in a heuristic fashion and the objective was to build an annotation system that starts with simple dictionary look-up and then evaluates and adjusts the results of the initial round of tagging using a sequential series of binary choices.[7] Multilingualiser incorporates a number of analytical tools in a stepwise fashion with several options for user-defined actions (see Figure 1).

[5] LiveCode Community, formerly RunRev, is available as an open-source application. A commercial license is required for distributing stand-alone applications. LiveCode source code can be compiled into OS X, PC and Linux distributions as well as into IOS and Android for mobile devices.

[6] Multilingualiser will be made available to the research community free of charge once the beta testing is over.

[7] The problem with training data is that for a task of this kind, the amount of manually tagged, gold standard data required would have been prohibitively large. As there was no training data available, it was impossible to use supervised learning models to develop an algorithm and consequently at least the current version of Multilingualiser is an unsupervised system.

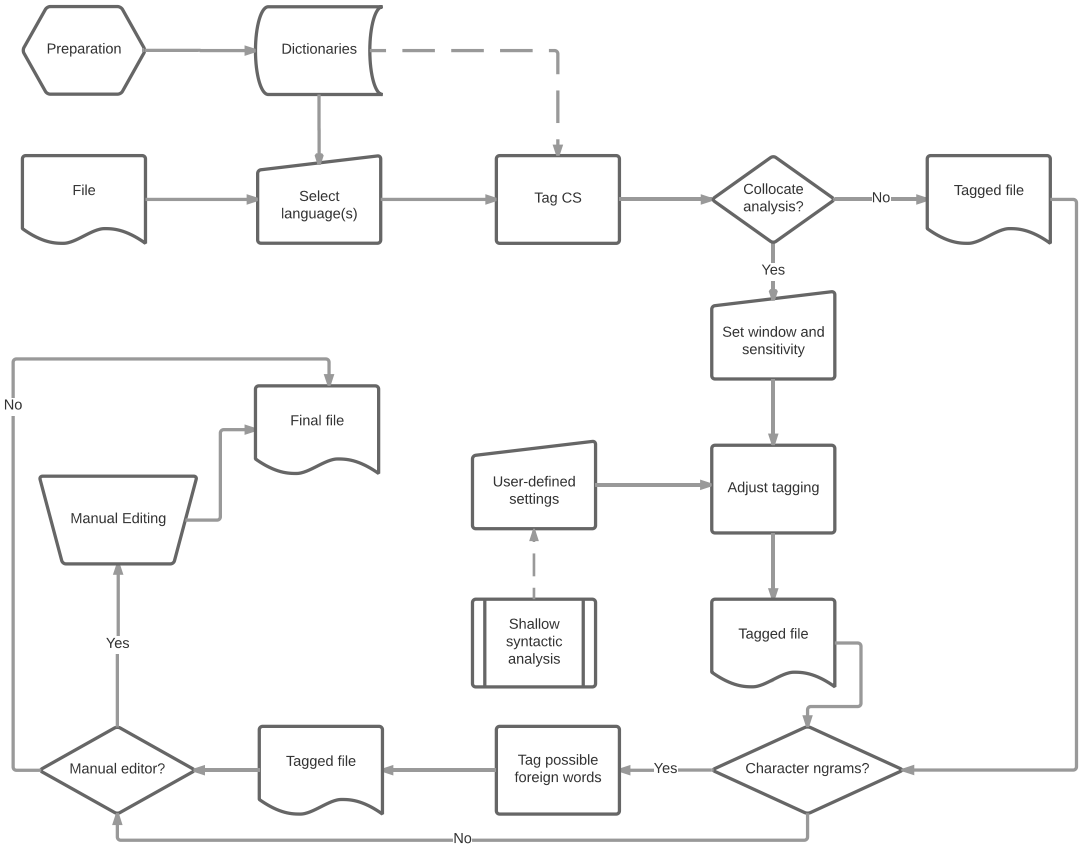

Figure 1: Flowchart of the Multilingualiser tagging process.

The main functionality of Multilingualiser begins with dictionary look-up, a computationally efficient approach of first-step pruning. Efficiency is a key word here, because the goal is to perform the analysis on large corpora of full-length texts using personal computers of standard specifications. For example, the longest text in CLMET3.0, Edward Gibbon's six-volume *The Decline and Fall of the Roman Empire* (1776–1788), contains 1.2 million tokens and nearly 35,000 unique types. The dictionary look-up requires at least one word list which contains either foreign words or, if preferable, words in the main language; Multilingualiser currently has built-in dictionaries for Latin, Italian, French and German, with separate word lists of function words in each language.

A Multilingualiser dictionary is a list of words which, if encountered in the text, will be tagged with the appropriate language tag. Each word is assigned a *foreignness score* (FS), a

numerical value that indicates the likelihood that the word is not an English word and, importantly, that it is a foreign word of a specific language; the scores are adjustable by the user.

In the development version of the Multilingualiser, used in the present study, FS values were first assigned in a heuristic fashion and then adjusted as the software development progressed. The next version of the software (in preparation) will include weighted scores based on new findings regarding foreign word collocates and entropy scores (Tyrkkö and Saarimäki forthcoming). The preparation of a new word list for Multilingualiser requires an initial round of processing in which a list of foreign words is cross-checked against a list of contemporary English words. All words not found in the English word list are given a high FS value, and all words found on both lists are given a low FS value. Because each FS score is language specific, the same word can appear in multiple dictionaries (see Figure 2). User-defined word lists can be used as well, which not only expands the range of languages but also allows the tagging of other types of items if so desired. Multilingualiser can tag texts one language at a time or incorporate any number of dictionaries in a single round of tagging.

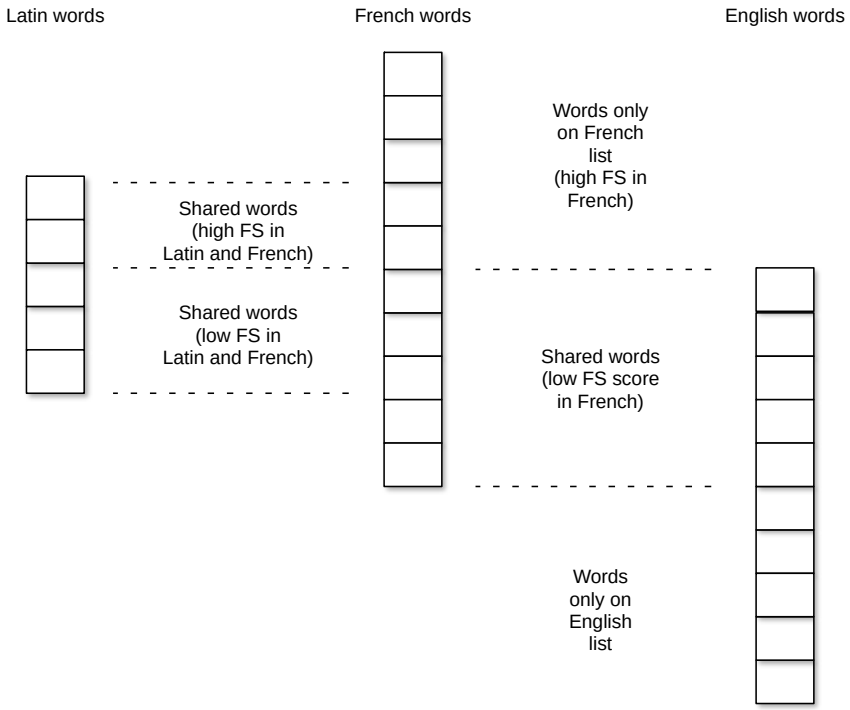

Figure 2: Dictionary cross-checking.

One of the particular challenges of tagging foreign words concerns diacritics. Although non-English diacritics are a good indicator that, for example, a word such as *molécule* is to be understood as a French rather than an English word, neither English authors, printers nor later digital editors can be relied upon to retain the original accents. Similarly, foreign words natively written with a diacritic, such as French *château*, may appear in English books without the accent. While this may seem like a minor problem, it creates considerable difficulties if the query engine interprets the accented and unaccented characters as different characters, reading *molécule* and *molecule*, or *château* and *chateau*, as two different words. The issue can be avoided either by including non-accented versions of all accented words in the foreign word list, which leads to a drop in precision, or by including them but assigning them a lower FS value.

Once the dictionary or dictionaries have been selected, the words in the source text are collected into a word list and the two (or more) lists are matched – the efficiency of this process can

be improved with a variety of optimization techniques such as recursive logarithmic searching and indexing.[8] The tool scores the matches (or misses, as the case may be) and their locations (by word number) as possibly foreign words. Because of the dictionaries, Multilingualiser is aware of which language or languages a word belongs to, and it is able to evaluate strings of more than one foreign language as highly unlikely, opting to retag items according to the most likely language.

Following the initial round of tagging, collocate analysis is run on suspected foreign words (see Figure 1). All the words within a user-defined span of words around each potential switch are analysed to determine whether a) any untagged words ought to be tagged, b) a tag ought to be changed or c) a tag ought to be removed. This process is repeated for each sequence of tagged words. Function words, or closed-class lexical items, are particularly useful here for disambiguating between closely related languages such as French, Italian and Spanish, and for ruling out visual diamorphs. Certain very high frequency diamorphs such as the English articles *a* and *an* can be left untagged if deemed best for processing efficiency; future versions of Multilingualiser will have a more fine-tuned range of options for managing these and similar items. At present, Multilingualiser identifies function words and it has a limited database of negative syntactic rules, that is, it knows which function words cannot form sequences. Sequence-initial and sequence-final words can also have special rules applied to them; for example, while capitalised words within sentences are generally proper nouns and thus ignored unless flanked by part of a longer sequence of foreign items, sentence initial capitalised words can – and usually should – be treated as regular words.

To revisit the previously discussed example (4), "no farther will I go than the duties of my *post oblige me*, and that honour, which to forfeit," Multilingualiser gives the highest likelihood to

[8] In *logarithmic searching*, instead of querying a long list of items one by one, the target array is recursively reduced by half until the remaining item is a match. In short, a word list is stored as an array variable and sorted into alphabetic order. The middle item of the array is tested against the query word to determine the half of the list in which the query word is. The process is repeated in a recursive manner until a match is found between the query word and an item on the list. Logarithmic searching speeds up the processing of long lists by a very large margin. *Indexing* is a method of pre-processing corpora in which each word type is treated as a unique item and the text position of each such occurrence as a value. Thus, to retrieve each instance of a particular type, we do not need to read through the entire text (or corpus), but instead we retrieve the item from the array and use the list of values to determine the position of each occurrence.

the interpretation that the words are English. Why? Although Multilingualiser recognizes *post* as a possible Latin word, the tool assigns a low Latin FS value to it because *post* is also an English word, and *oblige* and *me* get low French FS values for the same reason. If the analysis was stopped there, we would have a string of three words with low-value FS scores. However, the collocate analysis looks at a window around each word, defaulting at 5 words, to determine whether there are possible or definite foreign words within the span. Starting with *post*, Multilingualiser sees that there are no foreign words to the left. There are two words to the right, *oblige* and *me*, with the former getting a low FS value for French, and the latter scoring a low FS value for both. Multilingualiser thus recognises the sequence as one of two possibilities:

1. than_ENG the_ENG duties_ENG of_ENG my_ENG *post_LAT oblige_FRE me_LAT* ,_PNC
2. than_ENG the_ENG duties_ENG of_ENG my_ENG *post_LAT oblige_FRE me_FRE* ,_PNC

Because all the FS scores are low, Multilingualiser will by default approach them with prejudice. Since *post* has no right-side flanking items with a high FS value in the same language and *me* with a low FS value is not an immediately adjacent item, *post* is analysed to be a single word with a low FS score and Multilingualiser removes the tag. It proceeds to analyse *oblige*, which now has no foreign words to the left and *me* to the right. Since the two adjacent items have a low FS value in the same language, the analysis is strengthened that this may indeed be a foreign chunk.[9] Depending on the threshold value set by the user, the two words would be tagged as French with the option of including the low FS value in the tag. However, if the option of shallow syntactic analysis has been selected,[10] Multilingualiser can go further and determine that because the French object pronoun *me*

[9] If one of the two items had a strong FS value, the other item would receive a positive adjustment to its FS value. For example, *obligé d'en pleurer* would be analysed as a low FS value followed by two items with a high FS value, which would result in a positive adjustment to the FS value of *obligé*.

[10] The shallow syntactic analysis tool knows when function words require a word in the same language in the L1 or R1 position. The syntactic rule set has to be written specifically for each language.

requires a word in the R1 position and there is none, *me* is likely not to be French and the FS value is reduced.

For a more complex example, let us consider the following extract from *Red Pottage* (1899) by Mary Cholmondeley.

(6) Sincerity seems our only security against losing those who love us, the only cup in which those who are worth keeping will care to pledge us when youth is past. Rachel was not by *nature de celles qui se* jettent *dans l'amour comme dans un précipice*. But she shut her eyes, recommended her soul to God, and threw herself over.

The words identified by Multilingualiser as French are highlighted in italics, and two mistakes are underlined: *nature*, misidentified as a French word with a low FS value, and *jettent*, which was not in Multilingualiser's French dictionary and was consequently ignored. Furthermore, although *de*, *qui* and *un* are tagged as French, they are also potentially Italian (or Spanish or Portuguese, but Multilingualiser does not have the dictionaries at this time). So how does collocate analysis remedy the tagging? First, the words with multiple language tags are disambiguated by summing the overall FS values of the sequences; note that because *jettent* is not read as foreign, there are two chunks, one of five items and one of six. The French FS value is much higher than Italian, so Multilingualiser changes the Italian tags to French. Next, Multilingualiser identifies *jettent* as a lone untagged word in the middle of long strings of French words. It assigns *jettent* a low FS value tag for French, which allows it to be identified as French but does not suggest certainty. Finally, Multilingualiser notices the low FS value French tag for *nature* at the beginning of the long chunk. Depending on the user-defined settings, it can leave *nature* with a low-value FS tag or remove the tag; the latter choice guards against excessive lengthening of chunks but admittedly leads to some drop in precision.

Finally, two different character n-gram methods can be used to analyse words that do not appear in any dictionary (see Figure 1). Character n-grams can be used to analyse unusual or

distinct combinations of characters. Unlike Andersen (2011 and 2012), we do not implement a moving window but rather use a simpler method of evaluating initial trigrams against a dictionary of trigrams extracted from a large English word list,[11] and a series of language-specific word endings. Some of the words this method catches are simply rare English words, but an un-English combination of characters at the beginning or end of a word will flag it up as a potential multilingual element. Words identified with this method are then further evaluated in conjunction with their collocates, at which point only the most unusual words retain their foreign tags. This step works reasonably well because Late Modern English does not feature abbreviations and initialisms to the same extent Present-day English does; the same method would not work for contemporary data.

The current version of Multilingualiser has a bank of word endings in several languages and a list of unlikely word-initial trigrams. A selection of characteristic orthographic features is also included, such as the shortened forms of French reflexive pronouns *m'*, *t'* and *s'*, and the various contractions of prepositions and articles in Italian and Spanish. The character n-gram method has been successful in identifying some, but not all, foreign words in languages such as Arabic, Malay, Hindi, Hawaiian and Samoan. All in all, Multilingualiser returned chunks in 22 different foreign languages in CLMET3.0; all of these were checked, and most of the rarer languages were identified manually (see section 3.3). Several minor switches are also available for further refining the tagging process. For example, Multilingualiser can be told to ignore capitalised words within sentences – a simple form of named-entity recognition,[12] i.e. leaving out proper names[13] – and the collocate span can be adjusted manually by the user.

[11] We use the word list of the *Corpus of Historical American English* (COHA). Hapax legomena were pruned out to avoid including misspellings and very infrequently occurring words, such as foreign words. Another option would be to aggregate a list of English words from several large corpora, or to use a dictionary resource directly. Work is underway to incorporate HTOED data into Multilingualiser.

[12] Named entity recognition (NER) is a family of methodologies employed in information retrieval and text mining, typically used to identity names of persons, organisations and regions. A NER model will typically identify the names, including the proper segmentation of multi-part items, and classify them according to a training algorithm.

[13] A named-entity recogniser may be added to the release version of Multilingualiser.

Once the tagging is complete, the output can be saved locally in XML or the common underscore-tagged format. All foreign words are given the tag CS (for Code-Switch), followed by a distinct language code for each language in the dictionary (Lat, Fr, Ital and Ger) and a final "F" tag for function words. The XML version provides more choices, such as the addition of sequential word-id attributes, the preservation of pre-existing tagging such as part-of-speech data and an element for multi-word units. The FS value of each item can also be included as an attribute if necessary.

Multilingualiser also incorporates a built-in tag editor, which allows the user to examine a tagged text and to perform quick corrections. Typically, the tagger may have made a mistake with one particular word which is then repeated many times in the text. The user can remove such tags all at once, or one at a time. New words can also be tagged if necessary. The tag editor can perform simple chunking operations, saving sequences of tagged items with a contextual window in comma-separated values format for further analysis. The tag editor also allows the user to save all newly identified foreign words into the tagger's dictionary, which means that the tagging accuracy increases with each new file.

In terms of overall performance, assessing the accuracy of Multilingualiser in terms of traditional quantitative measures such as precision and recall is somewhat complicated. Because the tagger relies on language-specific dictionaries, the precision and recall rates depend greatly on the size and composition of the dictionaries used, as well as on a multitude of other factors such as spelling variants (especially with older texts), the retention of non-English characters such as accents, and the presence of items in languages for which no dictionary was available. By focusing on sequences of two or more words we were able to improve accuracy considerably, particularly because the longer the sequence is, the more likely it is to include items that appear in the dictionaries, such as function words and high frequency content words.[14] In practice, because our

[14] For a discussion of the frequency and collocation profiles of French and Latin items in CLMET3.0, see Tyrkkö and Saarimäki (forthcoming).

research design called for human evaluation of all instances, we opted in favour of higher recall – catching as many instances as possible, including false positives that were pruned out manually – at the cost of precision. Based on testing on human evaluated gold standard data, we believe we reached a precision of 0.69 and recall of 0.93, that is, perhaps one third of the chunks were false positives and a little less than 10% of all the chunks may have been missed. Although the performance of the tool is already sufficient to make it an asset in the analysis of multilingual practices, we wish to emphasize that our analyses of multilingual practices in Late Modern English were always based on human-evaluated data and never on the raw output of Multilingualiser.

There are also several shortcomings which will be addressed in the next version. One of the major drawbacks is that Multilingualiser is currently insensitive to dating, that is, that there is no mechanism for distinguishing between words that entered the language at different times. To remedy the situation, the next version of Multilingualiser will incorporate first datings from the database of the *Historical Thesaurus of the Oxford English Dictionary*,[15] allowing the user to cross-check the date of publication against the first recorded dating for each suspected foreign word. Thus, for example, *terra firma* could be given a high FS value for the first 50 or 100 years after the OED first dating of 1692, but a lower rating after that threshold. A further enhancement may see genre or register added as an analytical parameter, if the necessary metadata was available. This would mean that, for example, *an sich* would be tagged with a high likelihood of being foreign in fiction and drama, but it would get either a low FS value or perhaps none at all in scientific and philosophical texts. Common multi-word units, particularly ones that include visual diamorphs of common English words such as *a* and *an*, can also be included as a separate layer of analysis.

The use of frequency data is another potential area for development. At the simplest level this would mean using the frequencies of words in the corpus being analysed and flagging words that appear foreign but are also suspiciously frequent. A more sophisticated version would also use

[15] We wish to acknowledge our gratitude to the HTOED project at the University of Glasgow and to its current director Marc Alexander for granting us access to the database.

frequency data from period-appropriate megacorpora, thus at least partly addressing the well-documented challenges of using the OED datings; these include shortcomings with the balance and representativeness of sources from different centuries, differences in accuracy between the old and revised entries, and inconsistencies in the manner in which information on post-first dating frequencies is provided (see, e.g., Hoffmann 2004; Brewer 2007).

3.3 Post-processing: The human component

The discovery method applied to the CLMET3.0 data consisted of several rounds of testing, before the actual process began. Text extracts submitted for Multilingualizer were manually checked by humans for evidence of multilingual practices and these results were used to train the software further. Once the automated part of the process seemed to find a suitable balance of precision and recall, the first set of instances was retrieved.

In the beginning of the project, we started by retrieving all instances of possible multilingual expressions. However, the automatic recognition of one-word instances soon proved impractical. There are a great many English words that bear the morphology of their source even centuries after their nativization, with words such as *auditorium*, *inferno* or *confetti* appearing among the retrieved instances with great frequency. Identifying the extremely few, if any, cases where the usage could be interpreted as an unambiguous switch to another language would have required more time and effort than was possible to devote to the task.[16] The initial searches were run on files 259–270, and 68 out of 142 instances identified by Multilingualizer were classified as not being examples of multilingual practices. At this point Multilingualiser's collocate analysis module was adjusted to allow the automatic pruning of lone foreign words.

To make our reasoning behind excluding one-word instances more explicit, we will provide some examples of clear or fairly clear borrowings. Many of the terms, like *inferno* in (7) and

[16] The items were naturally included in the analysis whenever they appeared as part of a longer sequence of foreign words.

auditorium in (8), are used in the same text repeatedly. Others, like *cheval glasses* (9) or *guipure lace* (10), consist of a French borrowing combined with an English noun, and denote fashionable items of furniture or clothing, which are referred to by the borrowed term, which has not remained in use in English once the fashion itself passed. *Cheval-glass* is found in the OED (s.v. *cheval-glass*) as a compound, so it can be argued the initial part was already a borrowing in the nineteenth century. Similar terms referring to architecture, items of clothing and fashion more generally were e.g. *foyer*, *esplanade*, *rococo*, *domino*, *lorgnette* and *chic*. These are typically expressions which are still commonly used in English and refer to cases where both the concept and the term have been imported from another culture.

(7) They screamed if I didn't; and just as I was summoning the Almighty to attend to me a little in the middle of that *inferno*, out we came as innocent as a baby. There was another of these places just before getting into London. (Caine, *The Christian*)

(8) As she stepped down the stairs the curtain was drawn up, the *auditorium* was a void, the stage dark, save for a single gas jet that burned at the prompter's wing […] (Caine, *The Christian*)

(9) […] up a flight of stairs, and into a room of moderate size which had no window and no ventilation and contained three *cheval* glasses, a couch, four cane-bottom chairs, three small toilet tables with gas jets suspended over them, […] (Caine, *The Christian*)

(10) The bodice has windbag sleeves, formed of shawl pieces of *guipure* lace, and some lilies of the valley on the breast, finished with a waistband of heliotrope velvet […] (Caine, *The Christian*)

Obviously, there were also occasions where Multilingualiser recognised a word which was not a recent borrowing, but resembled, for one reason or another, words of another language. So, in (11) *hic* is not Latin but an onomatopoetic expression of the hiccups, and (12) is simply a case of mistaken identity, where *ral* in mimicked singing was misidentified as German. In (13)

Multilingualiser had tagged *riser* as French. There is a noun *riser* in French, but it is a fairly recent borrowing from English, and refers to a specific type of pipe on oil rigs (*Le Dictionnaire*, s.v. *riser*).

(11) Frayne. [Waving the past from him.] Yesterday– [with a slight hiccup] *hic*! [Turning away apologetically.] The heat in this room– [He walks away, as Sophy returns to Quex.] (Pinero, *The gay Lord Quex*)

(12) Jimmie. [To Lily, in a whisper.] Rat-tat, says the postman! [Catching hold of Roper and swinging him round.] La, ra, *ral*, la—! (Pinero, *The 'mind the paint' girl*)

(13) Ah, don't flatter yourself you're the only early *riser* in London. (Pinero, *The big drum*)

The decision to leave all one-word instances out of the data meant losing such genuine foreign-language expressions as can be seen in (14)–(16). The frequency of such accurate hits, however, was quite small compared to the number of imprecise search results.

(14) The man laughed to carry off his audacity. "Veil, you know vhat they say of us – agent from *agere*, 'to do,' and we're always 'doing.' (Caine, *The Christian*)

(15) But the women were either hopelessly *bourgeoises* or slightly *déclassée*. [Inspecting some of the pieces of *bric-à-brac* upon the table.] (Pinero, *The big drum*)

(16) […] he who excelled in it was known as "*Khâteb*," or Orator. (Haggard, *She*)

After these initial rounds of testing and fine-tuning the language-identification procedure, all 333 CLMET3.0 files were processed through Multilingualiser. The resulting raw tagged data was then manually checked in its entirety to determine whether the retrieved items were relevant and whether the language of the potential foreign-language sequences was correctly identified by the tool, as well as to verify their length in orthographic words. To identify languages that were unfamiliar to the manual checking team, various sources, including online dictionaries of possible languages and

annotated editions of the works in question, were used. At this stage, it was decided to exclude proper nouns such as names of books, paintings, streets and hymns (e.g. *Te Deum*) from further analysis. Since the data was being collected for a sociolinguistic study where the authors' multilingual practices would be related to language-external variables including their gender, educational background and occupation (see Nurmi et al. forthcoming), sequences that occurred in editorial notes or other parts that were clear later additions to the author's original text were also excluded, as far as their status could be determined from the context. A few instances of Scots such as the one in (17), tagged as foreign by Multilingualiser, were also left outside the analysed data.

(17) He instantly returned with some neighbours, and found the good woman seated amidst the advancing tide, which began to rise, with her lips ejaculating to her cummers, who she supposed were still pressing her to another cup, "*Nae ae drap mair*, I thank you kindly." We dined in family, and all well. (Scott, *The journal of Sir Walter Scott*)

The length of the potential switches needed to be confirmed as well, since Multilingualiser did not always tag all the foreign words in the sequence, or included a tag on an adjacent ambiguous word (cf. example 6).

Although the tagging process described above has a fairly high rate of recall for most foreign-language passages of two or more words in the material, Greek proved to be a challenge for Multilingualiser. There were 45 tagged sequences returned by the tool, found in 8 files, where the language was identified during the manual checking stage as Greek. This seemed very low in view of the role of Greek in British education during the eighteenth and nineteenth centuries covered by CLMET3.0, and analysis of the context of these and several other examples indicated that the corpus files in fact contained more Greek sequences which Multilingualiser had failed to identify as foreign.[17] A major reason for this is the composition of the corpus: its texts are taken from Project

[17] The version of Multilingualiser used in the study did not include a Greek dictionary, which means that the discovery of Greek words was entirely based on character n-grams. A dictionary of transliterated Greek lexis would naturally improve detection considerably, as would the tagging of all words that include characters outside the Latin alphabet as foreign.

Gutenberg and other online text archives, and were produced by numerous contributors using different principles and practices, including different character encoding schemes. Greek in the original source texts therefore appears in three basic forms in the corpus: indicated by an added textual marker where the Greek passage itself has been omitted, as in (18); transliterated (more or less accurately) into the Latin alphabet (19); or in the Greek alphabet (20a), sometimes simplified to exclude diacritics. The Greek in passages like (20a) typically shows up in the files as practically unrecognizable sequences of characters as a result of problems with the encoding when the source texts were converted into plain text files (20b). A fourth alternative is that the Greek content has simply been left out without any indication of the omission, and the only way to recover it would be to go through the original page by page. In theory, this applies to other foreign-language passages in the texts as well, but such elisions seem less likely in the case of foreign-language passages written in the Latin alphabet. The form in which foreign-language chunks appear in the corpus is thus related not only to the discourse practices of the texts' original authors and printers, but also to the choices of the modern-day transcribers of Project Gutenberg (for a discussion of alphabets and discourse practices, see Rütten, this volume). Of these four types, only the second one (19) could potentially be identified by Multilingualiser.

(18) […] for this clang too was of the imagination; preternatural; and it too walked in formless immeasurability, having made itself like to the Night (*Greek.*)! (Carlyle, *The French Revolution*) [In place of 'Greek.', the original reads νυκτὶ ἐοικώς]

(19) The belief of the *Greek eis oianos aristos amúnesthai perì pátrês*; the belief of the Roman that he was to trust in the gods of Borne, for those gods are stronger than all others; the belief of Cromwell's soldiery that […] (Bagehot, *Physics and Politics*) [The original reads εἷς οἰωνὸς ἄριστος ἀμύνεσθαι περὶ πάτρης]

(20a) […] most men are disguised by sobriety, and it is when they are drinking (as some old gentleman says in Athenæus), that men *ἑαυτοὺς ἐμφανίζουσιν οἵτινές εἰσιν*—display themselves in their true complexion of character, which surely is not disguising themselves. (De Quincey, *Confessions of an English Opium-Eater*)

(20b) […] most men are disguised by sobriety, and it is when they are drinking (as some old gentleman says in Athenæus), that men *ea?t??? eμfa?????s?? ??t??e? e?s??*—display themselves in their true complexion of character, which surely is not disguising themselves.

To counter these problems, we supplemented the results obtained using Multilingualiser with a separate round of searches for Greek in the CLMET3.0 files. In practice, this was carried out by first searching for the strings 'Greek' and 'Gr.', which returned all instances where the omission was marked using the name of the language or its abbreviation (as in example 18) as well as numerous cases of explicit flagging where the context included a word or phrase in Greek (as in 19), but inevitably also many hits that were irrelevant for the purposes of collecting examples of multilingual practices. The second round of searching consisted of searching for sequences with two question marks or a question mark followed by an alphabetic character. While admittedly inelegant, this search was successful in returning cases where the Greek script appears in the corpus files in mangled form. Although this dual method is by no means perfect in terms of either precision or recall, it raised the number of identified Greek chunks in the corpus six-fold to 270, spread across 39 files.

The manual evaluation of retrieved chunks also involved the analysis of switch type. The typological analysis deepens the analytical model, allowing us to investigate both the frequencies and types of switching in relation to a variety of sociolinguistic and text typological variables. This second part of the analysis is highly sensitive to both the immediate and the cultural context. Conventionalised switches, defined as recognisable foreign formulae such as greetings, words of common courtesy, academic phrases and so on, are multilingual passages which even a monolingual speaker of English might have in their idiolect (example 21). The use of such switches does not necessarily imply that the author knows the meaning of the words, though in most cases that is of course likely. Drawing the line between a conventionalised switch and a loan word can be particularly challenging. Pre-fabricated switches are quotes, proverbs and maxims, but also other

passages in a foreign language which can be originally attributed to someone other than the author although not in common use. All suspected switches of this type were extensively investigated using contemporary sources. The use of a pre-fabricated element suggests greater command of a foreign language than the use of conventionalised switching (example 22). Finally, as the label implies, free switches are, to the best of our knowledge, produced by the author and a demonstration of genuine command of the language in question (example 23).

(21) Now I assert, that whoever reasons after this manner, does *ipso facto* believe the actions of the will to arise from necessity […] (Hume, *A Treatise of Human Nature*)

(22) All which, from the words, *De gustibus non est disputandum*, and whatever else […] (Sterne, *The Life and Opinions of Tristram Shandy*)

(23) 'Mr Western,' answered the lady, 'you may say what you please, *je vous mesprise de tout mon coeur*. I shall not therefore be angry.' (Fielding, *The History of Tom Jones*)

Using this tripartite typological model, it is reasonably easy to identify pre-fabricated and free switches – though not always to say which is which – but deciding between conventionalised switches and loan words is much less straightforward. To give an example, the Latin phrase *terra firma* is attested in English from the early seventeenth century with the OED giving 1607 as the first dating. However, the current sense 'land as distinguished from the sea; dry or firm land' is first attested in 1692 and becomes reasonably frequent soon thereafter, featuring 41 times in CLMET3.0. Is *terra firma* a foreign expression in the eighteenth century? Or in the nineteenth? In example (24) from Disraeli's *Venetia*, published 1837, the expression is used without explanation. Similarly, *fille de chambre* is attested in English since 1673 and used by Sterne in 1768 without explication (example 25). Furthermore, dictionary evidence of first attestations needs to be evaluated carefully in the case of foreign-origin words, as the principles of inclusion may have varied.

(24) All four attendants immediately bowed, and extended their arms to assist this very great man; but Squire Mountmeadow, scarcely deigning to avail himself of their proffered assistance, and pausing on each step, looking around him with his long, lean, solemn visage, finally reached *terra firma* in safety, and slowly stretched his tall, ungainly figure. (Disraeli, *Venetia*)

(25) The fair *fille de chambre* came close up to the bureau where I was looking for a card […] (Sterne, *A Sentimental Journey through France and Italy*)

4 Concluding remarks

The study of historical multilingualism using large corpora is a new and challenging field in research that requires a combination of new methods and tools. Previous to the work discussed here, no research project has attempted to study historical multilingualism using a sufficient volume of primary data to allow truly frequency-based analysis of multilingual practices. By using a corpus some ten times larger than the largest corpora previously used in the study of historical multilingualism, such as the 1.5-million-word *Helsinki Corpus* (Pahta and Nurmi 2006) or the subsections of the *Corpora of Early English Correspondence* studied (usually approximately 0.5–1 million words in size, see, e.g., Nurmi and Pahta 2004; 2012 and Pahta and Nurmi 2011), we were able to reveal the pervasiveness and complexity of the phenomenon. This was one of the key challenges in the process, because prior information about exactly how, where and in what quantities foreign items occur in historical English texts of all registers and genres did not exist.

In this chapter we have reported on the methodological work of the *Multilingual Practices in the History of Written English* project,[18] outlining a set of semi-automatic methods of discovery and manual annotation that make it feasible to find, annotate and analyse foreign language content in medium-sized historical corpora such as the 34-million-word *Corpus of Late Modern English Texts 3.0*. To that end, we discussed the procedures built into Multilingualiser, our proprietary tagging tool, and the principles of the manual assessment and typological classification of foreign passages

[18] For the findings of the project, see, e.g., Nurmi et al. (forthcoming) and Tyrkkö and Nurmi (forthcoming) for frequency-based analysis of sociolinguistic and text typological predictors of multilingual usage, Tyrkkö and Saarimäki (forthcoming) for an analysis of collocational behaviour of foreign words in English texts, and Nurmi and Tyrkkö (in preparation) for a diachronic analysis of conventionalised foreign phrases in Early Modern and Late Modern English.

that we carried out on the computationally identified items. The discussion highlighted the need to approach multilingualism as a complex linguistic phenomenon that cannot be easily reduced to binary categories and requires careful analysis of the available data in its cultural and textual context.

References

Alex, Beatrice. 2005. "An Unsupervised System for Identifying English Inclusions in German Text." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), Student Research Workshop, 133–38.

Alex, Beatrice, Amit Dubey and Frank Keller. 2007. "Using Foreign Inclusion Detection to Improve Parsing Performance." In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 2007 (EMNLP-CoNLL 2008), 151–160. http://aclweb.org/anthology//D/D07/D07-1016.pdf. (accessed 18 November 2016).

Andersen, Gisle. 2011. "Corpora as Lexicographical Basis – The Case of Anglicisms in Norwegian." In *Methodological and Historical Dimensions of Corpus Linguistics* (Studies in Variation, Contacts and Change in English 6), edited by Paul Rayson, Sebastian Hoffmann and Geoffrey Leech. http://www.helsinki.fi/varieng/series/volumes/06/andersen/. (accessed 18 November 2016)

Andersen, Gisle. 2012. "Semi-automatic Approaches to Anglicism Detection in Norwegian Corpus Data." In *The Anglicization of European Lexis*, edited by Cristiano Furiassi, Virginia Pulcini and Félix Rodríguez Gonzáles, 111–30. Amsterdam: John Benjamins.

Androutsopoulos, Jannis. 2011. "English 'on Top': Discourse Functions of English Resources in the German Mediascape." *Sociolinguistic Studies* 6/2: 209–38.

Brewer, Charlotte. 2007. "Reporting Eighteenth-century Vocabulary in the OED." In *Words and Dictionaries from the British Isles in Historical Perspective*, edited by John Considine and Giovanni Iamartino, 109–35. Newcastle: Cambridge Scholars Publishing.

CLMET3.0 = *Corpus of Late Modern English 3.0*. Compiled by Hendrik De Smet, Hans-Jürgen Diller and Jukka Tyrkkö. https://perswww.kuleuven.be/~u0044428/.

Davidson, Mary Catherine. 2005. "Discourse Features of Code-switching in Legal Reports in Late Medieval England." In *Opening Windows on Texts and Discourses of the Past*, edited by Janne Skaffari, Matti Peikola, Ruth Carroll, Risto Hiltunen and Brita Wårvik, 343–51. Amsterdam and Philadelphia, PA: John Benjamins.

Dictionnaire, Le (s.a.) http://www.le-dictionnaire.com/.

Gardner-Chloros, Penelope. 2011. *Code-switching*. Cambridge: Cambridge University Press.

Gibbons, Edward. 1776–1788. *The History of the Decline and Fall of the Roman Empire*, vols. I–VI. London: Strahan and Cadell.

Grefenstette, Gregory. 1995, "Comparing Two Language Identification Schemes." In *Proceedings of Analisi Statistica dei Dati Testuali* (JADT), 263–68.

Heller, Monica. 1988. "Introduction." In *Codeswitching: Anthropological and Sociolinguistic Perspectives*, edited by Monica Heller, 1–24. Berlin, New York and Amsterdam: Mouton de Gruyter.

Hoffmann, Sebastian. 2004. "Using the OED Quotations Database as a Corpus – a Linguistic Appraisal." *ICAME Journal* 28: 17–30.

King, Levi, Sandra Kübler and Wallace Hooper. 2015. “Word-level Language Identification in *The Chymistry of Isaac Newton*.” *Digital Scholarship in the Humanities* 30/4: 532–40. Available online at DOI: <http://dx.doi.org/10.1093/llc/fqu032>. (advanced access 2 December 2014)

Kytölä, Samu. 2012. “Multilingual Web Discussion Forums: Theoretical, Practical and Methodological Issues.” In *Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse*, edited by Mark Sebba, Shahrzad Mahootian and Carla Jonsson, 106–27. New York and Abingdon: Routledge.

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Päivi Pahta, Heidi Koskela, Salla Lähdesmäki and Henna Jousmäki. 2011. *National Survey on the English Language in Finland: Uses, Meanings and Attitudes* (Studies in Variation, Contacts and Change in English 5). http://www.helsinki.fi/varieng/series/volumes/05/index.html. (accessed 18 November 2016)

Myers-Scotton, Carol. 1993. *Social Motivations for Codeswitching: Evidence from Africa*. Oxford: Clarendon Press.

Nurmi, Arja and Päivi Pahta. 2004. “Social Stratification and Patterns of Code-switching in Early English Letters.” *Multilingua* 23: 417–56.

Nurmi, Arja and Päivi Pahta. 2012. “Multilingual Practices in Women’s English Correspondence 1400–1800.” In *Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse*, edited by Mark Sebba, Shahrzad Mahootian and Carla Jonsson, 44–67. New York and London: Routledge.

Nurmi Arja and Jukka Tyrkkö. in preparation), “The Grey Border of Multilingual Practices: Conventionalised Items in Late Modern English” In *HiSoN2016 volume* (Studies in Variation, Contacts and Change in English), edited by Janne Skaffari, Päivi Pahta, Veera Saarimäki and Jukka Tyrkkö. Helsinki: Research Unit for Variation, Contacts, and Change in English.

Nurmi, Arja, Jukka Tyrkkö, Anna Petäjäniemi and Päivi Pahta. forthcoming. “The Social Embedding of Multilingual Practices in Late Modern English.” In *Multilingual Practices in Language History: New Perspectives*, edited by Päivi Pahta, Janne Skaffari and Laura Wright. (Language Contact and Bilingualism 15.) Berlin: de Gruyter.

OED = *Oxford English Dictionary Online*. http://www.oed.com/.

Pahta, Päivi. 2004. “Code-switching in Medieval Medical Writing.” In *Medical and Scientific Writing in Late Medieval English*, edited by Irma Taavitsainen and Päivi Pahta, 73–99. Cambridge: Cambridge University Press.

Pahta, Päivi. 2007. “On Code-switching in Early Modern English Medical Texts.” In *Tracing English through Time: Explorations in Language Variation in Honour of Herbert Schendl on the Occasion of his 65th Birthday*, edited by Ute Smit, Stefan Dollinger, Julia Huettner, Gunther Kaltenboeck and Ursula Lutzky, 259–72. Vienna: Braumüller.

Pahta, Päivi. 2011 “Code-switching in Early Modern English Medical Writing.” In *Medical Writing in Early Modern England*, edited by Irma Taavitsainen and Päivi Pahta, 115–34. Cambridge: Cambridge University Press.

Pahta, Päivi and Arja Nurmi. 2006. “Code-switching in the Helsinki Corpus: A Thousand Years of Multilingual Practices.” In *Medieval English and its Heritage*, edited by Nikolaus Ritt, Herbert Schendl, Christiane Dalton-Puffer and Dieter Kastovsky, 203–20. Frankfurt: Peter Lang.

Pahta, Päivi and Arja Nurmi. 2011. “Multilingual Discourse in the Domain of Religion in Medieval and Early Modern England: A Corpus Approach to Research on Historical Code-switching.” In *Code-switching in Early English* (Topics in English Linguistics 76), edited by Herbert Schendl and Laura Wright, 219–51. Berlin and Boston: Mouton de Gruyter.

Poplack, Shana. 1988. “Contrasting Patterns of Code-switching in Two Communities.” In *Codeswitching: Anthropological and Sociolinguistic Perspectives*, edited by Monica Heller, 215–44. Berlin, New York and Amsterdam: Mouton de Gruyter.

Schendl, Herbert. 1996. “Text Types and Code-switching in Medieval and Early Modern English.” *VIEWS* 5: 50–62.

Schendl, Herbert and Laura Wright, eds. 2011. *Code-switching in Early English*. Berlin and Boston: Mouton de Gruyter.

Stam, Nike and Tom de Schepper. forthcoming. “The Influence of Visual Diamorphs in Two Medieval Irish Corpora.” In *Multilingual Practices in Language History: New Perspectives* (Language Contact and Bilingualism 15), edited by Päivi Pahta, Janne Skaffari and Laura Wright. Berlin: de Gruyter.

Tyrkkö, Jukka and Arja Nurmi. forthcoming. “Analysing Multilingual Practices in Late Modern English: Parameter Selection and Recursive Partitioning in Focus.” In *Proceedings of ICAME36 (Language and Computers)* (Studies in Variation, Contacts and Change in English), edited by Sebastian Hoffmann and Andrea Sand. Helsinki: Research Unit for Variation, Contacts, and Change in English.

Tyrkkö, Jukka and Veera Saarimäki. forthcoming. “Profiling Clustered Foreign Words in English Historical Texts.” In *Patterns in Text: Corpus-driven Methods and Applications*, edited by Joanna Kopaczyk and Jukka Tyrkkö. Amsterdam: John Benjamins.

Voigts, Linda Ehrsam. 1989. “Scientific and Medical Books.” In *Book Production and Publishing in Britain 1375–1475*, edited by Jeremy Griffiths and Derek Pearsall, 345–402. Cambridge: Cambridge University Press.

Wright, Laura. 1999. “Mixed-language Business Writing: Five Hundred Years of Code-switching.” In *Language Change: Advances in Historical Sociolinguistics* (Trends in Linguistics: Studies and Monographs 114), edited by Ernst Håkon Jahr, 99–117. Berlin and New York: Mouton de Gruyter.

Wright, Laura. 2011. “On Variation in Medieval Mixed-language Business Writing.” In *Code-switching in Early English*, edited by Herbert Schendl and Laura Wright, 191–218. Berlin and Boston: Mouton de Gruyter.