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

How many languages are there in a monolingual corpus?

Arja Nurmi and Tanja Rütten

1 Introduction

The monolingual corpus as a monolithic, single-language database, representative of the language of likewise monolingual speakers or writers, is a tacit and probably only half-conscious, but convenient, invention by the corpus linguist. This is in line with the common societal assumptions of western societies about "one nation, one language" that rose in France during the revolution, dispersed over the nineteenth century in Europe and has dominated European thinking ever since. In linguistics this has inevitably resulted in an emphasis on the analysis of single languages, largely in isolation of each other. The notable exception from early on is research on language contact, examining the impact of one independent language system on the lexico-grammatical structure of another. However, not a single of the world's just over two hundred countries is monolingual (Deumert 2011, 262), and depending on our definition of bi- or multilingualism, it could be argued that the vast majority of the global population is in fact multilingual (see e.g. Edwards 2006, 7 or Li Wei 2007, 3–11). If we zoom in on Europe alone, a recent survey on Europeans and their languages carried out by the European Commission indicates that 54% of the population of EU member states meet the criterion for functional multilingualism, i.e. they are able to hold a conversation also in a language other than their mother tongue. To take an example from another corner of the world, the Australian census of 2006 lists 388 languages spoken in the homes of 16.8% of the population (Deumert 2011, 273). Surely, linguistic realities like these must have an impact on the authentic language data that corpus compilers store into their corpora. The question, then,

arises: Is multilingualism reflected in our corpora? If it is, how? And how do we as corpus linguists deal with it?

The question of how we define multilingualism is also relevant here (for the history of the concept, see e.g. Li Wei 2007). In this volume, multilingualism is seen, not as the traditional ideal of a balanced bilingual with a command of two languages that he or she has grown up with, but rather in terms of the speakers' linguistic resources and repertoires that originate in multiple languages, and their ability to apply those resources in their speech or writing. We see the potential for multilingualism both in individuals and in societies. Even if we do not necessarily agree with the position of Edwards (2006), who argues that modern speakers of English, who are familiar with such individual foreign-language words such as *tovarich* or expressions such as *Guten Tag*, are multilingual individuals, it is obvious that the definition of multilingualism should be inclusive of a range of abilities. Perhaps the most inclusive definition is given by Blommaert (2010, 102), according to whom

> [m]ultilingualism … should not be seen as a collection of 'languages' that the speaker controls, but rather as a complex of *specific* semiotic resources, some of which belong to a conventionally defined 'language', while others belong to another 'language'. The resources are concrete accents, language varieties, registers, genres, modalities such as writing – ways of using language in particular communicative settings and spheres of life, including the ideas people have about such ways of using, their language ideologies.

Even if we adopted a somewhat more restrictive outlook, remaining in the sphere of

different conventionally defined ‘languages’, we can safely say that monolingualism as a quality of either individuals or societies has always been a minority phenomenon. People throughout history have gained command of more than one language through education, professional contacts, personal interests, or migration – simply by virtue of living in a multilingual society and having to find ways to communicate with speakers of other languages. Even a very basic command of a language allows a speaker or writer to incorporate elements of it into their communication, i.e. to make use of their multilingual resources.

By way of experiment, if we turn our attention to a standard corpus of English, such as the *British National Corpus*, we can easily find many instances of multilingual practices that fit in an even stricter definition of multilingualism than that given by Edwards. The following examples were retrieved using random French, German and Latin phrases, and represent both informative (1, 3) and imaginative writing (2). Some searches reveal lengthier passages in another language, like example (1), which implies a considerable conversational fluency in the use of multilingual resources. Some hits occur in contexts that seem to prompt the use of the relevant language in the communicative situation, including reported conversations with speakers of other languages, as in example (2); in such contexts it is common to find several successive expressions in the same language. Again, some degree of competence in more than one language can be assumed. Interestingly, the search also reveals instances like example (3) where multilingualism reflected in the text does not rest on the speaker’s comprehension of multiple languages, which is a common criterion, used, for example, by Edwards (2006).

(1) After a while he returned, came over to me and, though I half expected a smack, said, ‘*Maintenant, il y a un nouvel relation entre nous. Maintenant nous serrons*

camarades.' We'd done it — (BNC: FS0 1727)

(2) The *fräulein* smiled and said, '*Auf wiedersehen*.' Karelius alone used the old Austrian farewell: '*Ich küsse die Hand*.' (BNC: B20 1488)

(3) What puzzles him, and us, is United's newly disencrusted coat of arms and its motto '*ex nihilo, nihil fit*.' I haven't the faintest what it means (BNC: K4T 9034)

As is apparent from examples (1)–(3), multilingual practices can also be seen as multivoiced practices, where quotations can represent the voice of someone other than the author (1, 3). Such quotations can also perform many of the same functions regardless of the language used, so that many English elements bear a resemblance to the French, German and Latin passages illustrated. Such quotative practices range from literary discussions and academic discourse conventions to language learning environments, where linguistic items from textbooks and teaching material are adopted and adapted to the linguistic repertoire of the learner. In both cases, speakers and writers make use of linguistic material that, in one sense, can be described as 'other than their own' and so produce a multivoiced text. While these multivoiced practices are not always multilingual (just as multilingual practices are not necessarily multivoiced), they bear a great deal of resemblance to multilingual practices, identified both in spoken language code-switching and written language data, and discussing them in this context will provide new insight into both phenomena. Figure 1 illustrates the relationship of the concepts of multilingual and multivoiced practices as we conceive them in this volume.

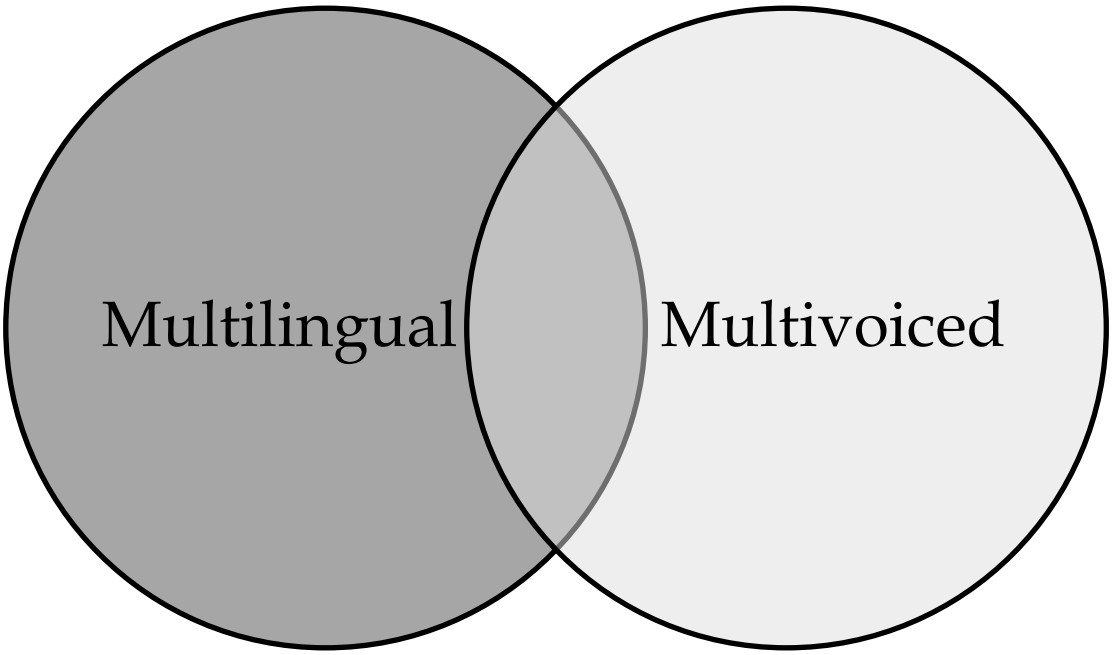

Figure 1. Multilingual and multivoiced practices

The combination of elements from more than one language, or voice, to a single communicative episode – whether a conversation or a text – thus appears much more common than is generally assumed, and may even be the rule rather than the exception. This point is supported in virtually all contributions collected in the volume at hand, from a historical as well as a present-day and cross-cultural perspective. It is also supported by e.g. Mair (2011), discussing the frequent use of Jamaican Creole in the spoken language of even educated Jamaican speakers in the ICE-Jamaica corpus. Mair further makes the point that in corpus-based studies of World Englishes multilingual contexts have been long ignored, and advocates for a more systematic study of multilingualism, both in interactive computer-mediated contexts and in spoken urban surroundings (2009, 436). On the other hand, recent research on some corpora compiled for analysing the history of English shows that multilingual practices are found in written texts from all historical periods (see e.g. Pahta and Nurmi 2011; see also Pahta et al. forthcoming). So it is time that we addressed the question of, firstly, just how many languages are there in what we

often assume are monolingual corpora of, say, English, and secondly, how can we compile corpora that better represent actual language use in contexts where standard English is just one of the varieties and languages in use?

This volume, then, brings together papers that investigate the presence of multilingual practices in supposedly monolingual corpora. The corpora discussed represent a broad range of Englishes and include present-day synchronic varieties of English as well as historical and diachronic perspectives. Contributions address the corpus compilers' views as well as the annotators' and users' perspectives. Viewpoints range from explicitly multilingual practices that are consciously taken into consideration in the compilation and annotation process to implicitly multi-voiced perspectives, where philological insight is used in unearthing multilingual practices in what superficially looks like a monolingual English corpus.

In the next section, we will briefly look at the sociological and language ideological underpinnings of the supposition of monolingualism in corpora (globalisation, superdiversity etc.). Section three presents the guiding questions for the volume and briefly reviews how individual contributions have answered them. Assessments range from the perspectives of research on multilingualism in the traditional sense of the concept to more innovative approaches, where the notion of multilingualism is extended to voices other than the author's and is thus halfway independent of the actual language that is used by the producer of the speech event. Section four rounds up this introduction by discussing ways to find, distinguish and describe non-English elements in 'monolingual corpora'.

2 Monolingualism – fact or fiction?

As mentioned above, monolinguals are a minority among the global population. Our focus in this volume is on English, hence we discuss the topic from that perspective, but many of the trends identified in English-speaking countries can also be seen elsewhere. In many countries different languages live side by side, are used in different registers and on different occasions. So in Tanzania, for example, speakers may have one native language they speak at home, while they are educated in Kiswahili, which is one of the lingua francas used also for e.g. business encounters. English plays a role in higher education and administration, and any number of other local languages may form a part of an individual speaker's linguistic repertoire (Melchers and Shaw 2011, 136). In terms of English world-wide, Meshtrie (2006, 482) goes as far as to claim that in these contexts monolingualism is "the marked case", while in the current globalising (or globalised) society, the "ideal speaker" encounters the need to draw on their linguistic resources in order to interact with people from all kinds of different backgrounds, whether in terms of solidarity or adversity, meeting as equals or negotiating power hierarchies. The "polyphony of codes/languages" can be seen as the native language of people in the context of New Englishes, but, in our view, more and more as the native language of people all over the world; the growing body of research on urban multilingualism and superdiversity provides ample evidence for this trend (see e.g. Blommaert and Rampton 2011, Creese and Blackledge forthcoming, Meyerhoff and Stanford 2015).

In addition to spoken interaction, multilingual practices are frequently in evidence in computer-mediated communication. It seems that there are still many hindrances for writing in non-prestige varieties, such as Jamaican Creole, in traditional media, unless it is for the purposes of folklore or quoting individual speakers. This has changed in e.g. diasporic online forums, where speakers make use of multiple languages and varieties to

construct their meanings. Mair and Pfänder (2013, 541) note that multilingual practices in their data are not a reflection of poor linguistic skills, but on the contrary they "are almost exclusively found with forum users who have full command of the normative varieties of the locally dominant languages and who thus use multilingual writing as an additional resource".

Is there any such thing as a monolingual speaker of English? If we consider the speakers of English in the world and their linguistic resources, it is evident that the only potentially monolingual group are the speakers of what Kachru (1985) calls "Inner Circle" Englishes: both the "Outer Circle", i.e. countries where English is spoken as a second language used in e.g. administrative and educational contexts, and the "Expanding Circle", i.e. the rest of the world where English is taught as a foreign language, are by definition contexts where speakers of English are largely multilingual. How monolingual then are the speakers of English in the Inner Circle?

Considering the situation of English-speaking countries, there are obviously autochthonous linguistic minorities in each and every one of them. (For details, see e.g. Melchers and Shaw 2011.) In the UK we find speakers of Welsh, Gaelic and Irish, in Ireland Irish is the national language beside English, in Canada apart from English and French there are speakers of First Nations and Inuit languages, and in the USA there are still many Native American and Alaska native languages. Similarly in Australia, there are speakers of Aboriginal languages and in New Zealand speakers of Māori. Many of these languages are endangered to varying degrees, although there are efforts to preserve them. In addition to the indigenous languages, there are many immigrant languages in each country, the smorgasbord of languages present in any community depending on the circumstances of migration. Immigrant languages may well have a long history as well,

considering e.g. the centuries of Spanish spoken in California. The communities of immigrant language speakers may be vitalised by new waves of migration, keeping the linguistic minorities from being assimilated. On the other hand, even long-standing linguistic minorities may well preserve some elements of their heritage language, even if they do not speak the language fluently any more. The numbers of European heritage-language speakers, especially Italian, German, Hungarian and French show a down-ward trend in US census data, but there are still approximately a million people resident in the United States who say they speak German at home (Ryan 2013). During the history of English, the waves of migrants, particularly Vikings and Normans, were slowly assimilated to the English-speaking population, but not without leaving their trace in the shape of English.

If we take one of the Inner Circle countries as an example, we can examine this situation in all its complexity. In the Irish census of 2011, 41.9% of respondents answered 'yes' to the question whether they could speak Irish (Central Statistics Office 2012). Given that all children learn both Irish and English at school, it could be argued that for a less strict interpretation of multilingualism, most people who have received their schooling in Ireland are multilinguals. In addition to the two national languages, schools also provide foreign language teaching in French, German, Spanish and Italian, which is in accordance with the EU language policy of everyone mastering two other languages in addition to their mother tongue (COM 2008). The 2011 census included for the first time also questions on other languages spoken at home, and 11% of residents reported they spoke a language other than English or Irish at home, the most common languages being Polish, French and Lithuanian. Of those speaking a foreign language at home, 6% answered they were not able to speak English at all. Given all this data, it could be argued

that the vast majority of Irish residents are multilingual to some extent.

As can be seen from the above example, not only do multilingual individuals gain their linguistic repertoires in a variety of ways but they also belong to a variety of different linguistic communities. In Ireland, for example, there are speakers of Irish living in the Gaeltacht area, where they encounter other native speakers of Irish and carry out many tasks related to their daily lives in Irish. At the same time, English is a part of their lives, as it is the overwhelmingly strong language of many areas of life. On the other hand, people who learn a foreign language at school (whether English in the Expanding Circle countries today or French or Latin in eighteenth-century England) are typically members of a far more loosely knit network of speakers.

Multilingual resources can be used for identity-work, marking membership in a linguistic group, as the Latino population in the United States does when they mix English and Spanish resources in their speech but also increasingly writing. Another type of identity-work is found in the Latin phrases found in the writings of well-educated people throughout the history of English. There the writers can indicate their own membership in the community of educated people but they can also build bridges towards their readership, marking them as members of the same educated elite. The less educated would have had fewer linguistic resources in the range of multilingualism, but even they had access to e.g. Latin as the language of religion, engaging in both multilingual and multivoiced practices when referring to the teachings of the church.

3 Challenging the myth of monolingual corpora

In corpus linguistics, increasing the size, but not necessarily the quality of the database has been one of the major goals for a long time. Ever bigger databases, resulting in

automatic, web-crawling ‘corpora’ (e.g. in the case of GloWbE) seemed to be on the top of corpus linguists’ wish lists, and for good reasons. At the same time, it should be noted that the “small and tidy” and “big and messy” approaches of corpus compilation and annotation both have their merits (see e.g. Mair 2006 for a discussion of this). While it is true that corpus enhancement along the lines of automatic tagging and parsing has always been a major branch of corpus linguistic activity, too, the question of how to deal with non-English elements in English language corpora has seen considerably less scholarly activity. Size does matter, for an assessment of multilingual practices as well as for nearly everything else, but in order to identify multilingual practices in the first place, improved annotation is essential, too. And in order to improve annotation schemata, a sound idea of what constitutes a multilingual element is, of course, a necessary prerequisite.

When discussing the annotation of multilingual elements, the question of language boundaries comes up. At times, language users clearly flag their other language elements and their switches from one into another (Poplack 1987). In speech this can take place for example through repetition or metalinguistic commentary, but also pauses, hesitation and the mention of the language switched into. In writing, similar tendencies can be seen, and in English historical writings, for example, flagging can take the form of explicit labelling (*that is in Latin*), or in the case of foreign-language elements the reader might not be able to understand easily, the introduction of intratextual translation or support in English, often highlighted through either verbal (or, *i.e.*, *that is to say*) or visual cues (parentheses, italics, underlining) (Nurmi and Skaffari 2016). Elements accompanied with flagging elements like these are easily recognised as evidence of multilingual practices. Once they are identified in the text, they are also relatively straight-forward to annotate. There are, however, also times when speakers and writers deal with their linguistic output in a way

that has been described as translanguaging (see e.g. Otheguy, García and Reid 2015). On these occasions, writers do not pay attention to the boundaries between languages, but rather treat all their linguistic resources as one pool of features to draw from in order to communicate their meaning. These instances may also be occupying the grey area between borrowing and multilingual practices, as they may fluidly use both domesticated and original spelling, for example. In present-day spoken Finnish the English adverbial *about* (in the sense 'approximately') is frequently used. When it is written, the written form can follow standard English spelling (6), but can also reflect the domesticated spoken form (e.g. *öbaut* or *abaut* in 4 and 5), even in quality newspapers such as the *Helsingin Sanomat*.

(4) "Viime vuoden kesäkuusta tämän vuoden kesäkuuhun työllisten määrä on kasvanut 33 000:lla. Jos pystyttäisiin pitämään tällainen trendi vuoteen 2019 asti, oltaisiin 72 prosentin työllisyysasteessa, *öbaut*", Sipilä sanoo. (*Helsingin Sanomat* 12 August, 2016)
'"From June last year to June this year the number of the employed has risen by 33,000. If we could maintain a trend like this until 2019, we would be at an employment rate of about 72%", says [Prime Minister] Sipilä.'

(5) Asun tossa *abaut* sadan metrin päässä Evästiellä. (*Helsingin Sanomat* 4 November, 1999)
'I live there about a hundred meters away, in Evästie.'

(6) Se oli *about* vartti kun äijiltä lähti lapasesta. (@JethroRostedt on Twitter 4 March, 2015)
'It was about a quarter of an hour before the guys lost it.'

Considering that all spelling and pronunciation variants from Standard English to variously domesticated Finnish perform the same function in the texts and maintain the English meaning, trying to pigeon-hole these expressions into separate categories of code-switching/code-mixing and borrowing would be not only futile but counterproductive in terms of speakers' linguistic production. This also presents a dilemma for corpus coding. How to deal with such hybrid elements in-between languages? This is an issue that is

particularly of interest for corpora of more informal language, whether spoken or written, but since these elements tend to find their way even to the quality newspapers, initially through interviews and columns, trying to decide on a particular moment as a cut-off point is difficult without a good understanding of the current status of any individual linguistic element.

With these issues in mind, the contributions in this volume address the following questions:

1. From a corpus compiler's view:
What to do with multilingual texts and elements, when compiling a monolingual corpus? What are the criteria for inclusion and exclusion in sampling? How does representativeness play into these choices?

2. From a corpus annotator's view:
How to annotate foreign-language passages in a corpus? Should they be given a text-level coding, and if so, how detailed? In case of linguistic annotation, how should foreign-language elements be dealt with?

3. From a corpus user's view:
How can we study multilingual practices in monolingual corpora? How do we approach a corpus, if the foreign-language elements have not been annotated? How do we deal with questions of representativeness, if the corpus compilers have not in any way indicated their choices with regard to multilingual elements? What kinds of results on multilingual practices can be gained when studying multilingual practices in supposedly monolingual corpora?

For obvious reasons, these three views are often intertwined. For example, the question of how we can study multilingual practices in a (seemingly monolingual) corpus depends, of course, on the amount of annotation with which the respective corpus is equipped. In a similar way, the question how detailed an annotation schema should be depends, amongst other things, on the multilingual practices of the population from which this

sample stems.

Consequently, all contributions in this volume consider most, if not all, of the above questions, but place emphasis on different aspects. Research perspectives range from Postcolonial and World Englishes over a range of non-native and learner Englishes to historical stages of the language. The corpora described in the individual contributions discuss explicitly multilingual practices in the traditional sense of the concept as well as more opaque multi-lingual and multi-voiced discourse practices.

Of the papers that discuss explicit multilingual practices in seemingly monolingual corpora, the opening paper of this volume by Lange reviews how multilingual practices are documented in the various postcolonial components of the *International Corpus of English* (ICE). In particular, Lange evaluates ICE-India from both a corpus user's and a corpus compiler's perspective, and discusses building a more balanced corpus of Indian English with view of the multiple native languages influencing the Englishes spoken on the subcontinent.

In a similarly explicit multilingual context, Onysko and Degani discuss the selection of texts and informants for a corpus of mono- and bilingual native speakers of New Zealand English, with the concomitant problems of coding both background information and text level variation. They also place emphasis on the question how cultural meaning can be explored by corpus-linguistic means, provided the respective annotation schema systematically accounts for the diversity of multilingual elements in the corpus.

Besides these obvious multilingual contexts provided by postcolonial varieties of English, the myth of monolingual practices also extends to corpora compiled to study non-native and learner Englishes, and English as a lingua franca. These lines are pursued

in the three subsequent contributions. First, Laitinen brings to table a discussion of annotating the multilingual elements in advanced non-native corpora of English, when the languages used range from majority languages to traditional minority languages and immigrant languages.

An explicit learner perspective is pursued in the contribution by Callies and Wiemeyer, who introduce the *Corpus of Academic Learner English* (CALE). Callies and Wiemeyer discuss various approaches to annotating multilingualism and transfer in learner corpora and describe developing an annotation practice for multilingual elements. Their contribution is complemented by Kreyer's paper, towards the end of the volume, who discusses multivoiced practices in learner Englishes, which turn out to be much more implicit than the phenomena introduced in Callies and Wiemeyer.

Hynninen, Pietikäinen and Vetchinnikova approach English as both a spoken and written lingua franca in academic and private contexts (ELFA and WrELFA corpora of academic spoken and written ELF). Their focus is on a discussion of the appearance and functions of multilingual practices in English as a Lingua Franca. In all three cases, multilingual practices occur quite explicitly in the data but are dealt with in various ways in both the compilation process and in the way in which the data were approached to conduct research.

From a diachronic perspective, explicit multilingual practices are discussed in the contributions by Kohnen, Rütten, and Tyrkkö, Nurmi and Tuominen. Kohnen presents ideas for building a corpus of commonplace books – strikingly similar to Laitinen's present-day corpora of non-native Englishes in their presentation of often complete texts in one language in a multilingual compilation or environment. From a research-oriented perspective, Kohnen also explores basic questions of language choice in the genre of

commonplace books.

Rütten introduces the annotation schema developed for the *Corpus of English Religious Prose* against the background of the long-standing history of multilingual practices in the religious domain. In addition, she describes multivoiced practices in the domain, which may or may not be multilingual, and illustrates how these can be dealt with in the corpus architecture and basic annotation.

By contrast, Tyrkkö et al. take a turn on (semi-)automated processes of identifying multilingual elements in an unannotated corpus. In addition to describing software designed to reliably identify, annotate and analyse foreign language elements in a historical English corpus, the *Corpus of Late Modern English 3.0* (CLMET3), Tyrkkö et al. emphasise that multilingual practices cannot be reduced to binary distinctions, e.g. foreign/English, native/non-native English, as is often conveniently done. Instead, they show how textual and cultural context feed into an assessment of multilingual practices.

Against the background of these explicit multilingual practices in synchronic and diachronic corpus linguistics, Kreyer and Kaunisto introduce more opaque, multivoiced practices. These appear much more implicitly in corpora, but are strikingly similar to multilingual practices (see also section 2). Both Kreyer and Kaunisto, and also Rütten in her discussion of the "invisible hand", offer different approaches to multivoiced texts, discussing intertextual elements that represent another speaker's or writer's voice in a text, whether multi- or monolingual. Of these papers, Kreyer seemingly takes the notion of multilingualism in corpora to its very limits. Turning to learner corpora, Kreyer discovers the extent to which learner texts are mere copies of source material in the *Marburg Corpus of Intermediate Learner English* (MILE). In fact, being multivoiced in this sense, such learner productions resemble multilingual practices to a considerable

extent. Consequently, Kreyer discusses the types of mark-up needed to detect such multivoiced practices and provides an illustrative analysis of intermediate learner English in MILE. Kaunisto takes a corpus user's perspective and conducts a philological study of Samuel Taylor Coleridge's *Biographia Literaria*, which is one of the files contained in the *Corpus of Late Modern English* (CLMET3), but lacks any form of multilingual annotation. He shows how severe the influence of multivoiced interference can be even on high frequency items such as personal pronouns.

All contributions agree that various languages, in varying proportions, appear alongside with English in the "English" corpora which are investigated in this volume. Depending on their respective research paradigms, contributors offer various courses of action for this situation. This highlights the fact that we may be well advised to rethink our understanding of corpora as monolingual language data repositories. Also, we need to address the question how to find and interpret non-English elements.

4 Tracing multilingual practices in supposedly monolingual corpora

How does one find, distinguish and describe foreign language elements in both, corpora that do and corpora that do not flag non-English elements as such? In theory, there are two general routes one may wish to take here: automatic and manual identification. In the real world, the task is usually a combination of both.

In the present volume, Tyrkkö et al. present a semi-automatic approach, introducing software that identifies non-English elements with considerable precision. Rütten presents a corpus design that integrates multilingual, and to a lesser extent also multivoiced, practices into the architecture of the corpus from the start. At the other extreme, the contributions by Kaunisto and Kohnen proceed from purely philological

points of departure, identifying multilingual elements with the help of scholarly editions and informed philological knowledge about context (text production, text reception, circulation etc.). While both approaches will successfully identify non-English elements, only the latter is able to spot multivoiced elements. The identification of multivoiced elements is something that might be of interest in corpus research, and could be at least partly automated in the future, since familiar quotations could be identified using electronic text repositories, and other flags for multivoiced elements could be identified (at least the use of quotation marks and quotative phrases like *he/she says* and *according to*).

However, this is a vital challenge in research on multilingual practices, as is pointed out in several contributions. Hynninen et al. show that even though corpus compilers may flag a linguistic structure as non-English, this need not necessarily be the case for the speakers in the actual speech events. Hynninen et al. look at how code-switches are flagged in discourse and they see a noteworthy discrepancy between explicitly flagged code-switches by the speakers and annotation schemata by compilers that only distinguish English from foreign elements. While the foreign-tag marks non-English elements, these tags may say very little as to how code-switches were perceived by the actual speakers. This, of course, has implications for the assessment of the level of competence of non-native English speakers and brings in another facet of multilingualism that may need attendance in the annotation schema.

Along the same lines, Kreyer contrasts materials and task descriptions from the English language learning environments with students' textual productions. His findings indicate that even advanced learners show one third of their collocations as originating from the materials/task descriptions. Again, this not only has implications for the

assessment of language competence and idiomaticity, but points to yet another issue to be taken into consideration in annotating supposedly monolingual material.

Far from being able to resolve these matters within the two covers of this book, we hope that bringing these issues into focus will help to rethink the widely accepted notion of 'the monolingual corpus' and to be able to better fine-tune into text samples, knowing that much can be expected that is not the voice, or language, of the author.

References

Blommaert, Jan. 2010. *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press.

Blommaert, Jan and Ben Rampton. 2011. "Language and Superdiversity." *Diversities* 13/2: 1–22.

Central Statistics Office. 2012. *This is Ireland. Highlights from Census 2011, Part 1*. Dublin: Stationary Office.

COM. 2008. "Multilingualism: An Asset for Europe and a Shared Commitment." Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels: Commission of the European Communities.

Creese, Angela and Adrian Blackledge, eds. forthcoming. *The Routledge Handbook of Language and Superdiversity*. London and New York: Routledge.

Deumert, Ana. 2011. "Multilingualism." In *The Cambridge Handbook of Sociolinguistics*, edited by Rajend Mesthrie, 261–282. Cambridge: Cambridge University Press.

Edwards, John. 2006. "Foundations of Bilingualism. In *The Handbook of Bilingualism*, edited by Teij K. Bhatia and William C. Ritchie, 7–31. Oxford: Blackwell.

European Commission. 2012. *Europeans and their Languages*. Special Eurobarometer 386. http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf.

Kachru, Braj B. 1985. "Standards, Codification, and Sociolinguistic Realism: The English Language in the Outer Circle." In *English in the World: Teaching and Learning the Language and the Literature,* edited by Randolph Quirk and H.G. Widdowson, 11–30. Cambridge: Cambridge University Press.

Li Wei. 2007. "Dimensions of Bilingualism." In *The Bilingualism Reader,* edited by Li Wei, 3–22. 2nd edition. London: Routledge.

Mair, Christian. 2006. "Tracking Ongoing Grammatical Change and Recent Diversification in Present-day Standard English: The Complementary Role of Small and Large Corpora." In *The Changing Face of Corpus Linguistics,* edited by Antoinette Renouf and Andrew Kehoe, 355–376. Amsterdam: Rodopi.

Mair, Christian. 2009. "Corpus Linguistics Meets Sociolinguistics: Studying Educated Spoken Usage in Jamaica on the Basis of the International Corpus of English (ICE)." In *World Englishes: Problems, Properties, Prospects,* edited by Lucia

Siebers and Thomas Hoffmann, 39–60. Amsterdam: Benjamins.
Mair, Christian. 2011. “Corpora and the New Englishes: Using the ‘Corpus of Cyber-Jamaican’ (CCJ) to Explore Research Perspectives for the Future.” In *A Taste for Corpora: In honour of Sylviane Granger*, edited by Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin and Magalie Paquot, 209–236. Amsterdam: Benjamins.
Mair, Christian and Stefan Pfänder. 2013. “Vernacular and Multilingual Writing in Mediated Spaces: Web Forums for Post-colonial Communities of Practice.” In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives,* edited by Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi, 529–556. Berlin and New York: de Gruyter.
Melchers, Gunnel and Philip Shaw. 2011. *World Englishes*. 2nd edition. London: Hodder Education.
Meshtrie, Rajend. 2006. “Society and Language: Overview.” In *Encyclopedia of Language and Linguistics*, Vol. 11, edited by Keith Brown, 472–484. Amsterdam: Elsevier.
Meyerhoff, Miriam and James N. Stanford. 2015. “‘Tings Change, All Tings Change’: The Changing Face of Sociolinguistics with a Global Perspective.” In *Globalising Sociolinguistics: Challenging and Expanding Theory*, edited by Dick Smakman and Patrick Heinrich, 1–15. London and New York: Routledge.
Nurmi, Arja and Janne Skaffari. 2016. “*Whiche is in Englisshe tong* –Managing Latin in English.” Paper presented at the International Conference on English Historical Linguistics (ICEHL-19), Essen, August 22–26.
Otheguy, Ricardo, Ofelia García and Wallis Reid. 2015. “Clarifying Translanguaging and Deconstructing Named Languages: A Perspective from Linguistics.” *Applied Linguistics Review* 6/3: 281–307.
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,* edited by Herbert Schendl and Laura Wright, 219–251. Berlin: Mouton de Gruyter.
Pahta, Päivi, Janne Skaffari and Laura Wright, eds. forthcoming. *Multilingual Practices in Language History: New Perspectives*. Berlin and New York: de Gruyter.
Poplack, Shana. 1987. “Contrasting Patterns of Code-switching in Two Communities.” In *Aspects of Multilingualism,* edited by Erling Wande, Jan Anward, Bengt Nordberg, Lars Steensland and Mats Thelander, 51–77. Uppsala: Borgströms.
Ryan, Camille. 2013. *Language Use in the United States: 2011*. American Community Survey Reports. Washington, DC: U.S. Department of Commerce.