Use logbooks and find the original meaning of "representativeness"

Andrea Ágnes Reményi
Research Institute for Linguistics,
Hungarian Academy of Sciences
Budapest, Hungary
(remenyi@nytud.hu)

Debates about overall problems in general synchronic corpus design seem to have settled. Yet, solutions whether and how to pre-structure one's target population and corpus are still based either on practical considerations of comparability or on intuitive proportions, partly due to an exclusively textual view of representativeness, and partly to the fact that designers deny the possibility to estimate the relative distribution of texts among media, genres, registers, etc. of a language. In this paper first general problems are tackled again: what do 'representativeness', 'balance' and 'influentialness' mean? In the second part a relatively simple and cost-efficient method to estimate those distributions is described, and results of a two-step pilot study are analysed. Finally I will suggest how both textual and demographic representativeness can be controlled in a modular corpus.

General synchronic corpora wish to grasp the totality of a language in some sense. Electronic corpora support reliable quantitative studies only if the sample selected from this totality represents the totality as fully as possible. Leech (1991) states that "a corpus is 'representative' in the sense that findings based on an analysis of it can be generalised to the language as a whole or a specified part of it" (as cited by Kennedy 1998: 62). In my view a sample can be called 'representative' only if it aims to reproduce the statistical variance of the population to the highest possible degree – in terms of the distribution of text types and of linguistic features (cf. Biber 1993: 243), but not excluding demographic factors.

1 Theoretical considerations

The requirement of representativeness poses a major difficulty in corpus design, as designers are supposed to find valid bases to delimit the concept of "every text in the given language", i.e. the target population, on the one hand, and to find the most reliable sampling methods to select from that, on the other.

1.1 Statistical sampling.

Branches of social science applying statistical sampling and inference most often employ either random sampling or stratified random sampling methods. In the former case every member of the target population has an equal chance to appear in the sample. When the latter method is applied, researchers previously establish a few basic categories considered by the researchers/the research community/the research paradigm to be structuring the target population in some essential way (for example sex, age, schooling, location; medium, genre, domain, etc.), and random sampling is achieved only within these 'strata'.

1.2 Stratified random sampling.

When defining the target population of a general synchronic corpus, designers must also start by deciding which sampling method to apply, that is, whether or not they should pre-structure the population by a classification of text types or language users, a structure that is maintained in the sample. Johansson (1980:26), for example, supports a textual stratification over simple random sampling, stating that "the true 'representativeness' of the LOB Corpus arises from the deliberate attempt to include relevant categories and subcategories of texts rather than blind statistical choice. Random sampling simply ensured that, within the stated guidelines, the selection of individual texts was free of the conscious or unconscious influence of personal taste or preference."

Note, however, that this is not only a methodological, but also a conceptual problem, as 'strata' of the population are pre-defined by the researchers, necessarily reflecting what *they consider* to be the most essential structuring factors of that population: textual, demographic or other factors.[1]

[1] Apart from this, the most obvious advantage of stratified random sampling over simple random sampling is that it requires a smaller sample.

1.3 What are the units of observation?

The basis of this meta-theoretical issue is the double nature of the *units of observation* in corpus linguistics. While in sociology the units of observation are mostly unambiguous (individual human beings), this is not so in corpus design. Should language users (text producers and receivers) or texts (the products of language use) be chosen as the units of observation? While census or similar statistical data about the totality of language users of a country are easily accessible in most countries, no similar methods have been developed to estimate either the totality of the target population of texts or the distribution of text types within it. Thus in text-based sampling the basis of the composition and proportioning of 'strata', i.e. text types, is statistically unjustifiable. To my knowledge, most megacorpora (Brown, LOB, Cobuild, BNC, ICE, Longman, etc.) are structured according to medium, register, genre, domain, discourse function and/or subject in a way that proportions of these text types are determined by the practical consideration of *balance*, that is, by including comparable amounts of texts within subcategories.

On the other hand, corpora organised by demographic proportions would not support the criterion of 'sample variability matching population variability' as far as text types are concerned. Biber is right in stating that "a corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing. [...] Such a corpus would permit summary descriptive statistics for the entire language represented by the corpus. These kinds of generalisations, however, are typically not of interest for linguistic research" (1993: 247). *Some* linguistic research, however, e.g. sociolinguistics or L2-lexicography and textbook methodology, may find interest in a demographically structured corpus.

To sum up, the problems of 'representativeness' are mostly due to the double nature of the unit of observation in corpus design: either the diversity of language users, or that of text types is eclipsed.

1.4 A combination of the two types of units of observation.

Among the above mentioned corpora, the spoken component of the British National Corpus (BNC) tried to solve this dilemma, pairing text-based and demographic sampling methods. In the text-based ('context-governed') part of the corpus both major and minor *a priori* categories were proportioned to be balanced (four equal-sized contextually based categories were established, each divided into 40 per cent monologue and 60 percent dialogue) (Burnard 1995: 23), consonant with the proportioning of the written component. In the demographically sampled part of the BNC's spoken component 124 individuals were recruited based on random location sampling, asked to carry a tape recorder and to record all their conversations for 2-7 days. "Recruits were chosen in such a way as to make sure there were equal numbers of men and women, approximately equal numbers from each age group, and equal numbers from each social grouping." (BNC Online 1997.) Note that stratified sampling was performed only in terms of location, but not in terms of the other basic demographic variables, e.g. sex, age, and socio-economic group,[2] i.e. the proportions of age groups or socio-economic groups did not follow those in the UK population. Thus, while the BNC has a complex structure of 'strata', stratified sampling according to these basic demographic variables within none of them is achieved.

The conversation subcorpus of the Longman Spoken and Written English Corpus was also sampled on demographic bases: "a set of informants was identified to represent the range of English speakers in the country (UK or USA) across age, sex, social group, and regional spread" (Biber et al. 1999: 29), but the informants may not have been randomly selected for that corpus, either.

1.5 Is it possible to estimate the distribution?

Designers usually dismiss the possibility to take the daily distribution of spoken and written medium, domain, genre, register, discourse function or subject variation into account (e.g. Biber 1993:247, Burnard 1995: 20, Kennedy 1998: 63). There are direct methods developed to obtain the production and indirect methods to estimate the reception of published written and internet texts (book and periodical publication lists; best seller, library lending and periodical circulation statistics; internet search engines and click-on per internet page measures, respectively). Direct quantitative methods to assess the reception of published written texts, and the production and reception of unpublished written texts have not been found, and neither are there objective measures to define the

[2] Tamás Váradi directed my attention to this point.

target population of the spoken medium available. Kennedy writes: "No one knows what proportion of the words produced in a language on any given day are spoken or written. Individually, speech makes up a greater proportion than does writing of the language most of us receive or produce on a typical day. However, […] a broadcast conversation on radio or television will reach many more ears than a commercial encounter involving just a customer and a salesperson. Within a written corpus, balance is equally intractable. […] How to get a balance between the few writers and speakers who are prestigious and the great majority of text producer and speakers who have no special claim to fame is not simple" (1998: 63). Biber also stresses the factor of influentialness: "proportional samples are representative only in that they accurately reflect the relative numerical frequencies of registers in a language — they provide no representation of relative importance that is not numerical" (Biber 1993:247-8; the author shifts the problem to culture studies, similarly to Sinclair 1991: 13). These authors seem to be mixing up the concept of 'representativeness' with the concept of an external factor, that of 'influentialness'.

1.6 'Influentialness'

As I have stated, a general corpus should be as diverse as possible to fulfil the statistical axiom that every effort is to be made to reproduce the variance of the population in the sample. But the non-controlled early introduction of influentialness rules out the possibility of *the study of variable influentialness* of texts later, in the analysing stage (as, for example, in a sociolinguistic investigation about the factors of certain sayings becoming proverbs, while others not).

An example from sociology may illustrate my point. Network analysts can undoubtedly suppose that certain members of the population are more influential, and their sampling methods support the study of the effect of this influentialness as a variable. It would be a mistake to collect data only of the 'influential' members of society the same way as it would have been a mistaken move by early demographers and sociologists to include only those individuals (i.e. well-educated upper-/middle-class men over, say, 40 years of age) in their samples, but, fortunately, the tradition never developed in *that* discipline.

2 The logbook method

How can we estimate the daily distribution of medium, domain, genre, register, discourse function and/or subject variation in the population of a given language satisfactorily? Data collection based on tabular format *logbooks* filled in by a demographically representative sample of (adult) speakers yields a reliable picture of these distributions. The logbook consists of sheets of tables informants are asked to carry along their daily activities, and to fill in rows of cells whenever they use language, excluding self-talk and meta-notes ('filling in the logbook'). Information is to be collected about:

- the duration of the activity (in minutes or seconds)
- whether the informant is the producer, a receiver or a third party (role-changes indicated in the same row)
- the approximate number of participants (if known)
- brief demographic details of other participants (if assessable)
- the medium (spoken or written)
- the setting (home, office, street, shop, church, etc.)
- genre (based on a list of possible genres, but extendible).

Other factors (e.g. the domain, or the subject or aim of language use) can be added, controlled the same way by extendible lists given in the short guide that is fastened to the booklet of the fill-in tables. A more detailed manual must be given to each informant, who must also be shortly trained.

Data-collection per informant must take two to seven days, at least two days when the informants' activities are characteristically different, e.g. a weekday and a weekend day.

To develop the logbook method, I have conducted two pilot studies.

2.1 Pilot study 1

First a small pilot study was carried out to figure out if the logbook method with the structure described above was at all feasible. Two individuals took notes while travelling on public transport for a few days. The procedure was executable, and, apart from notes on conversations with acquaintances and strangers, greetings, chance remarks to strangers (e.g. '*Sorry!*'), newspaper and book reading, the

study yielded the yet unacknowledged genres of browsing mega-posters, eavesdropping on others' conversation, reading shopping lists, etc.

2.2 Pilot study 2

The seven informants participating in this data collection (five females and two males) filled in the logbook tables for a full day whenever they used Hungarian, starting either early morning and finishing when going to sleep, or starting at any time on one day, and finishing at the same time the following day. Data was produced about three weekdays and three weekend days (the seventh informant started on Sunday, and finished on Monday).

The logbook-booklet included several sheets of A4 tables (see the format below), fastened together with a short guide. (For the guide and a list of abbreviations, see the Appendix.) Informants were also verbally instructed how to fill in the tables, what to watch for, and were informed about the aim of the study.

Informant:..... Date: ..... Started at (hour, minute): .....

duration	You (P/R/3)	No. of Rs[3]	the others	wr/sp	setting	genre	(subject)	(aim)

A new row in the table was to be filled in every time when either the number of participants, the medium, the setting or the genre changed, when, for example, a new participant joined or left the interaction. As the subject was not of primary research importance, subjects within a type of linguistic activity were allowed to be listed within one cell. When two types of language use were either alternating or happening parallelly, the informants were asked to connect the two rows describing them with brackets. E.g. if the informant was reading and at the same time listening to other people talking; or if spontaneous conversation was regularly interrupted by interesting news on TV.

When I collected the logbooks, I asked the informants to comment on it or the task: they clarified ambiguous details, and gave plenty of comments, both about the pitfalls of the logbook and their experiences. They were all surprised to realise that so much time was spent on so many different types of linguistic activities. One of them even called it consciousness-raising task, because now she realised that her whole life was being spent almost on nothing else but giving and receiving verbal, visual and other signals.[4]

2.3 Results

Table 1 shows some quantitative results: the overall duration[5] of linguistic activity by the seven informants in minutes (mean: 622.86 minutes, standard dev.: 240.2), also broken down by medium. As it can be expected, a majority of all these people's language use was in the spoken medium, though with a highly varying proportion.

2.3.1 Spoken activities

The proportion of longer spontaneous face-to-face conversations (LSFC) with family, acquaintances or strangers were calculated for each informant (see Table 1). The remaining time within the spoken medium was spent on telephone conversations, fleeting interactions (greetings, saying *Sorry!*, paying in a shop), watching movie films or TV-programmes, listening to radio programmes or to others' conversations and loudspeaker announcements while travelling on public transport, mixed classroom activities, etc.

[3] "Number of participants" would have been a better label.

[4] All my informants were extremely helpful, for which I am indebted.

[5] In the case of parallel or alternating linguistic activities, the duration spent on each activity was calculated separately.

age	sex	time of week	sum (min.)	written (min.)*		spoken (min.)*		LSFC**	
40	female	weekend day	739	95	13%	644	87%	503	78%
39	female	weekday	1058	509	48%	549	52%	451	82%
24	male	weekend day	736	175	24%	561	76%	540	96%
46	female	weekend day	653	196	30%	457	70%	455	99%
12	female	weekend-weekday	502	114	23%	388	77%	125	32%
31	female	weekday	400[6]	192	48%	208	52%	141	68%
36	male	weekday	272[7]	32	12%	240	82%	227	95%

Table 1. Duration of Hungarian linguistic activity by informant in pilot study 2 (in minutes); * = percentage given in proportion to sum; ** = percentage of longer spontaneous face-to-face conversations (LSFC) given in proportion to spoken activity

2.3.2 Written activities

Informants' activities included reading newspapers (various types and columns), reading books (non-fiction: popular science, professional, school textbook; fiction), browsing megaposters, reading a map, writing and reading e-mails, taking notes while reading, checking one's own paper manuscript, browsing a library catalogue, leafing through books and magazines in a library or bookshop, reading posters, billboard advertisements, tourist signs, bus signs, streetname plates and sign-boards in the street, reading out a bedtime story to a child, entering data into a mobile phone, writing a school-test and homework, filling in a library request card, using a Hungarian-language word-processor while correcting an English text. The bedtime tale, a regular activity for parents with young children, is a genre that was difficult to classify, because it is a mixture of a written text read aloud and spontaneous conversation.

2.4 Informant sensitivisation

A disadvantage of the logbook method may be that is not robust enough to counterbalance differences in informant sensitivity. To raise the method's reliability level, informants should be carefully trained to be able to consciously check all their activities. The pilot study indicated that certain communicative activity types were not acknowledged, or were not broken down in sufficient detail, by some informants. For example, classroom interaction is a multi-genre linguistic activity. Similarly, several genres mix in a morning TV or radio programme: news-reading (the informant is a recipient of a text written down to be read aloud), interviews (the informant is a third party of others' spontaneous conversation) and commercials (the informant can be either or both). The detailed logbook manual must also include informative definitions of setting, genre, subject, etc. with plenty of supporting examples.

2.5 Advantages

This method, while not yielding language-use data, provides a statistically reliable picture of the daily distribution of text types in the demographic population, which can be exploited in the compilation of a demographically oriented corpus. While methods to determine the target population of published written and internet texts are available, the logbook method offers a solution to assess the target population of spoken and unpublished written texts. It also makes a comparison with an existing general corpus based on genre, register, or other classification possible. Moreover, new genres can be found, and the connection between the number of recipients and 'relative importance' or 'influentialness' of texts can be studied. The logbook method seems simple and relatively cost-efficient, thus it can be based on a large sample of speakers.

2.6 Modular corpora

One may ask: "Do we need representative corpora at all? Monitor corpora are collected with a primary emphasis on the quantity of texts, and if the texts are well-documented, the user can decide on the proportions." The problem is that most users (lexicographers, syntacticians, etc.) are not informed about problems of corpus design, and have no tools to assess populations but their own

[6] She reported having spent most of her day writing a paper in English.
[7] He reported having been writing an English language computer programme all day.

intuitions. It is the corpus designer's and the sociolinguist's task to assess the textual and demographic population not only for reference, but also for monitor corpora.

If texts in a monitor corpus are well-documented as far as the relevant categories ('strata') are concerned, a user-interface picking texts according to given criteria can produce any composition of texts, with the possible control over other criteria. Thus, with the help of such an interface either a balanced build-up of texts can be composed (e.g. for a multi-genre comparative analysis of a syntactic feature), or a demographically proportional one — based on the results of a survey using the logbook method (e.g. for a sociolinguistic study on 'influentialness', or for another study on the spread pattern of politically correct phrases). The end-user's preferred composition could also be set (e.g. including only the coded spoken texts for a prosodic analysis, or texts produced in a given year for a study on new coinages), or the totality of the modular corpus's texts could be used, as well. When the interface has randomly picked the texts according to the given criteria, it must give summary statistics about the composition of texts, so that the analyst could change the criteria if the composed corpus is too small. Misbalance in the original monitor corpus (causing also end-user corpora to be too small) can be adjusted by cyclical fine-tuning (Biber 1993, Váradi 1998) by including missing text types. Such an interface may help to exploit the possibilities of monitor corpora.

References

Biber D 1993 Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243-257.

Biber D, Johansson S, Leech G, Conrad S, Finegan E 1999 *Longman grammar of spoken and written English*. Harlow, Longman.

BNC Online (1997) http://info.ox.ac.uk/bnc/what/spok_design.html

Burnard L 1995 *The BNC Handbook*. Oxford, Oxford University Press.

Johansson S 1980 The LOB Corpus of British English texts: Presentation and comments. *ALLC Journal* 1 (1): 25-36.

Kennedy G 1998 *Introduction to corpus linguistics.* London-New York, Longman.

Leech G 1991 The state of the art in corpus linguistics. In Aijmer K, Altenberg B (eds), *English corpus linguistics: Studies in honour of Jan Svartvik*. London, Longman, pp 8-29.

Sinclair J 1991 *Corpus, concordance, collocation*. Oxford, Oxford University Press.

Váradi T 1998 *Nyelv és korpusz: a reprezentativitás a korpusznyelvészetben (Language and corpus: representativeness in corpus linguistics).* Manuscript. Budapest, Research Institute for Linguistics.

Appendix: The logbook guide for pilot study 2.

Carry the booklet with this guide for a full day wherever you go,
and fill in the lines every time you use the Hungarian language.
(Except when thinking/talking to yourself or when filling in the tables.)

duration: of the linguistic activity (how many minutes/seconds did it take?)
You (P/R/3): were you a **P**roducer/**R**eceiver/third party?
No. of Rs: an approximate number of **R**eceivers
others: brief demographic details of other participants — if known
(sex, age, profession, residence: Budapest/town/village)
wr/sp: written/spoken (or other: e.g. written text read aloud)
setting: e.g. shopping, cinema, conference, religious service, etc.
genre: OTHER GENRES CAN BE NAMED, TOO!

- two- or several-party personal conversation
(e.g. with family, acquaintances, strangers in the street, in a shop, in the doctor's office, etc., or listening to others' conversation as a third party)
- two- or several-party personal conversation
(e.g. with family, acquaintances, strangers)
- fleeting interaction with acquaintance, stranger (e.g. only greetings, saying sorry, asking for a journal at the news-stand, etc.)
- giving presentation in front of a present audience/listening to one
- listening to a loudspeaker announcement
- work meeting
- radio/TV: news, interview, debate, sports commentary
- radio/TV/tape recorder/CD: music with text (if you listen to text)
- radio/TV/movie film, theatre play
- radio/TV: commercial
- reading: newspaper, magazine, internet
(column: news report, editorial, fiction, advertisement, etc.)
- reading: book (fiction, science, law, etc.)
- reading: short message (letter, shopping list, etc.)
- reading: megaposter, poster, announcement, brand names on buildings or clothes, sign-board
- writing:: short message (letter, shopping list, etc.)
- writing: creation of longer text (diary, paper)

Optional:
subject
aim: e.g. playing, giving or receiving information/orders, amusement, killing time, etc.
If two types of language use are either alternating or happening parallelly, connect the two rows describing them with brackets.

Thanks a lot for your effort!