Magali Paquot
Stefan Th. Gries *Editors*

A Practical Handbook of Corpus Linguistics

MOREMEDIA

Springer

A Practical Handbook of Corpus Linguistics

Magali Paquot • Stefan Th. Gries
Editors

A Practical Handbook of Corpus Linguistics

Springer

Editors

Magali Paquot
FNRS
Centre for English Corpus Linguistics,
Language and Communication Institute
UCLouvain
Louvain-la-Neuve, Belgium

Stefan Th. Gries
Department of Linguistics
University of California
Santa Barbara, CA, USA

Justus Liebig University Giessen
Giessen, Germany

ISBN 978-3-030-46215-4 ISBN 978-3-030-46216-1 (eBook)
https://doi.org/10.1007/978-3-030-46216-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introduction

Corpus linguistics is “a whole system of methods and principles” (McEnery et al. 2006: 7f) that can be applied to answer research questions related to language use and variation in a wide variety of domains of linguistic enquiry and beyond. Over the last decades, it has been “among the fastest-growing methodological disciplines in linguistics” (Gries 2015: 93) and is now also developing as a key methodology in the humanities and social sciences.

The tasks of corpus linguists are manifold and complex. They can be grouped into minimally three different, though of course interrelated, areas:

- **Corpus design**, which requires knowledge about *corpus compilation* (e.g., the notions of sampling and/or representativeness), data processing for *corpus annotation* (e.g., tagging, lemmatizing, parsing), and *corpus architecture* (e.g., representing corpus data in a maximally useful way); then, once there is a corpus,
- **Corpus searching/processing**, which requires knowledge of, ideally, *general data processing* (e.g., file management, dealing with different annotation formats, using regular expressions to define character strings that lead to good search results) as well as *corpus query tools and methods* (from off-the-shelf tools to programming) to address the specificities of various data types (e.g., time alignment of spoken and multimodal corpora or bitext alignment of parallel corpora); then, once there are results from a corpus,
- **Statistical analysis** to get the most out of the corpus data, which requires knowledge of *statistical data wrangling/processing* (e.g., establishing subgroups in data or determining whether transformations are necessary), *statistical analysis techniques* (e.g., significance testing or alternative approaches, regression modeling, or exploratory data analysis), and *visualization* (e.g., representing the results of complex statistical analysis in ways non-quantitatively minded readers can understand).

This handbook aims to address all these areas with contributions by many of their leading experts, to be a comprehensive practical resource for junior and more senior corpus linguists, and to represent the whole research cycle from corpus creation, method, and analyses to reporting results for publication. It is divided into six parts. In Part I, the first three chapters focus on **corpus design** and address issues related to corpus compilation, corpus annotation, and corpus architecture. Part II deals with **corpus methods**: Chapters 4–9 provide an overview of the most commonly used methods to extract linguistic and frequency information from corpora (frequency lists, keywords lists, dispersion measures, co-occurrence frequencies and concordances) as well as an introduction on the added value of programming skills in corpus linguistics. Chapters 10–16 in Part III review different **corpus types** (diachronic corpora, spoken corpora, parallel corpora, learner corpora, child language corpora, web corpora, and multimodal corpora), with each chapter focusing on the specific methodological challenges associated with the analysis of each type of corpora.

Parts IV–VI aim to offer a user-friendly introduction to the variety of statistical techniques that have been used, or have started to be used, more extensively in corpus linguistics. As each chapter under Part IV–VI uses R for explaining and exemplifying the statistics, Part IV starts with an introductory chapter on how to use R for descriptive statistics and visualization. Chapters 18 and 19 focus on **exploratory techniques**, i.e., cluster analysis and the multidimensional exploratory approaches of correspondence analysis, multiple correspondence analysis, principal component analysis, and exploratory factor analysis. Part V focuses on **hypothesis-testing** (classical monofactorial tests, fixed-effects regression modeling, mixed-effects regression modeling, generalized additive mixed models, bootstrapping techniques and conditional inference trees and random forests). It is important to note that the chapters on mixed effects regression modeling and generalized additive mixed models in particular are primarily meant for readers to get a grasp of what these techniques are (more and more corpus linguistic papers rely on such methods and it is important for corpus linguists to understand the current literature). However, a single chapter can of course not provide all that is required to get started with statistics of such a level of complexity. If the reader is interested to know more and wants to use these statistics for their own purposes, they will necessarily need to read more on the topic.

Part VI aims to pull everything together by providing guidelines for how to write a corpus linguistic paper and how to meta-analyze corpus linguistic research.

Chapters in Parts IV and V as well as Chaps. 7, 9 and 27 come with online additional material (R code with datasets).

It is our hope that this handbook will serve to help students and colleagues expand their methodological toolbox. We certainly learned a lot while editing this volume!

Louvain-la-Neuve, Belgium Magali Paquot

Santa Barbara, CA, USA Stefan Th. Gries

References

Gries, S. T. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics, 16*(1), 93–117.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London/New York: Routledge.

Contents

Part I Corpus Design

Part II Corpus methods

Part III Corpus types

Part IV Exploring Your Data

Part V Hypothesis-Testing

Part VI Pulling Everything Together

Part I
Corpus Design

Chapter 1
Corpus Compilation

Annelie Ädel

Abstract This chapter deals with the fundamentals of corpus compilation, approached from a practical perspective. The topics covered follow the key phases of corpus compilation, starting with the initial considerations of representativeness and balance. Next, issues in collecting corpus data are covered, including ethics and metadata. Technical aspects involving formatting and annotation are then presented, followed by suggestions for sharing the corpus with others. Corpus comparison is also discussed, as it merits some reflection when a corpus is created. To further illustrate key concepts and exemplify the varying roles of the corpus in specific research projects, two sample studies are presented. The chapter closes with a brief consideration of future directions in corpus compilation, focusing on the importance of compensating for the inevitable loss of complex information and taking the increasingly multimodal nature of discourse as a case in point.

1.1 Introduction

Given that linguistics is descriptive at its core, many linguists study how language is used based on some linguistic sample. Finding the right material to use as the basis for a study is a key aspect of the research process: we are expected to use material that is appropriate for answering our research questions, and not make claims that go beyond what is supported by the material. This chapter covers the basics of compiling linguistic material in the form of a corpus. Corpus compilation involves "designing a corpus, collecting texts, encoding the corpus, assembling and storing the relevant metadata, marking up the texts where necessary and possibly adding linguistic annotation" (McEnery and Hardie 2012:241). In the process of putting together linguistic data in a corpus, researchers need to make a series of decisions at different steps. The process is described in a general way in this chapter, while more

A. Ädel (✉)
Dalarna University, Falun, Sweden
e-mail: annelie.adel@du.se

M. Paquot, S. Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*,
https://doi.org/10.1007/978-3-030-46216-1_1

in-depth discussion relating to the compilation of specific types of corpora follows in Chaps. 10–16. Specifics on corpus annotation and corpus architecture follow in Chaps. 2 and 3, respectively.

1.2 Fundamentals

1.2.1 Representativeness

The most basic question to consider when compiling a corpus involves representativeness: what type of speakers/variety/discourse is the corpus meant to represent? In many of the well-known corpora of English, the ambition has been to cover a general and very common type of discourse (such as ‘conversation in a variety of English’) or a very large population (such as ‘second-language learners of English’). However, such a comprehensive aim is beyond the scope for most researchers and should be reserved for large groups of researchers with plenty of resources at their disposal (see e.g. Aston and Burnard (1998) for discussions on how the *British National Corpus* was designed, or Johansson et al. (1978) on the *Lancaster-Oslo/Bergen Corpus*). In small-scale projects, the aims regarding representativeness need to be more modest by comparison, for example with a focus on a specialized type of discourse used by a relatively restricted group of speakers.

The general sense of the word ‘sample’ is simply a text or a text extract, but in its more specific and statistical sense it refers to “a group of cases taken from a population that will, hopefully, represent that population such that findings from the sample can be generalised to the population” (McEnery and Hardie 2012:250).[1] The aim in compiling a corpus is that it should be a maximally representative—in practice, this translates into *acceptably* representative—sample of a population of language users, a language variety, or a type of discourse. In most linguistic studies, we have to make do with studying merely a sample of the language use, or variety, as a whole. It is only in rare cases, and when the research question is quite delimited, that it is possible to collect all of the linguistic production of the population or type of discourse we are interested in. As an example, it may be possible for a researcher in Languages for Specific Purposes to retrieve all of the emails sent and received in a large company to use as a basis for studying the typical features of this specific type of communication in that company.

The corpus builder needs to consider very carefully how to collect samples that maximally represent the target discourse or population. One of the ways of selecting material for a corpus is by stratified sampling, where the hierarchical structure (or ‘strata’) of the population is determined in advance. For example, a researcher

[1] Samples in the sense ‘text extracts’ are occasionally used in corpora to avoid having one type of text dominate, just because it happens to be long. There are many arguments for using complete texts, however. See e.g. Douglas (2003) and Sinclair (2005).

who is interested in spoken workplace discourse could document demographic information about speakers' job titles and ages and whether interactions involve peers or managers/subordinates, and then include in the corpus a predetermined proportion of texts from each category. In the detailed sampling process, it is decided exactly what texts or text chunks to include.

There is a range of possible considerations to take in deciding about sampling procedures for a corpus, one of which concerns to what extent to organize the overall design around text production or text reception. For illustration, this is what the compilers of the *British National Corpus* (Aston and Burnard 1998:28) concluded with respect to the written part of the corpus:

> In selecting texts for inclusion in the corpus, account was taken of both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have a wide distribution. Thus, having chosen to sample such things as popular novels, or technical writing, best-seller lists and library circulation statistics were consulted to select particular examples of them.

A concept that is intrinsically related to representativeness is balance, which has to do with the proportions of the different samples included in the corpus. In a balanced corpus, "the relative sizes of each of [the subsections] have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled" (McEnery and Hardie 2012:239). In the case of 'conversation in a variety of English', the researcher would need a principled way of deciding what proportions to include, for example, of conversations among friends versus among strangers, or unplanned versus preplanned conversations (an interview is an example of the latter), or conversations from institutional/public/private-personal settings, and so on. Such decisions could be based on some assessment of how commonly these different configurations occur or of their relative importance (however this may be defined). Balancing decisions could even be based on comparability with some other corpus: for example, in a diachronic corpus of English (cf. Chap. 10) fiction writing may be deliberately overrepresented and religious writing underrepresented in earlier periods to allow for easier comparison to present-day English.

The notions of representativeness and balance are scalar and vague (see e.g. Leech 2007), so there are no hard and fast rules for achieving representativeness and balance in a corpus. The first step is to map out the available types of discourse, in order to find useful categorizations of the different ways of communicating used in the target community. The point that the most important consideration in corpus compilation is "a thorough definition of the target population" which is able to describe the "different situations, purposes, and functions of text in a speech community" was made by Biber (1993:244–245) in a classic piece on representativeness in corpus design. Added to this are "decisions concerning the method of sampling" (Biber 1993:244), as the next step is to find some principled way of representing these different ways of communicating. For some of the early standard corpora, this was done by drawing on classifications from library science, where there is a long tradition of cataloguing written publications. For example, a list of the collection of books and periodicals in the Brown University Library and the Providence Athenaeum was used as a sampling frame for the pioneering *Brown corpus*, aiming to represent written American English in general (published in 1961); see Francis and Kucera (1979).[2] Using stratified random sampling, a one-million word corpus was produced, consisting of 500 texts including 2,000 words each.

However, if the available types of discourse are not already classified in some reliable way, as in the case of spoken language, it means that the corpus builder will have to dedicate a great deal of time to researching the characteristics of the target discourse in order to develop valid and acceptable selection criteria. Douglas (2003) describes this type of situation and includes a useful discussion about the collection of *The Scottish Corpus of Texts and Speech*.

With a definition of representativeness as the extent to which a corpus reflects "the full range of variability in a population" (Biber 1993:243), it has been suggested that representativeness can be assessed by the degree to which it captures not only the range of text types in a language (external criteria), but also the range of linguistic distributions in a language (internal criteria). Since different linguistic features—vocabulary, grammar, lexicogrammar—vary in frequency and are distributed differently "within texts, across texts, across text types" (ibid.), the corpus should make possible analysis of such distributions. In fact, Biber (1993) suggests a cyclical method for corpus compilation, including as key components theoretical analysis of relevant text types (which is always primary) and empirical investigation of the distributions of linguistic features. However, few corpus projects have attempted this.

The literature on corpus design sometimes contrasts 'principled' ways of building a corpus to 'opportunistic' ones. An opportunistic corpus is said to "represent nothing more nor less than the data that it was possible to gather for a specific task", with no attempts made "to adhere to a rigorous sampling frame" (McEnery and Hardie 2012:11). It is, however, very difficult not to include some element of opportunism in corpus design, as we do not have boundless resources. This is especially true of single-person MA or PhD projects, where time constraints may present a major issue. What is absolutely not negotiable, however, is that the criteria for selecting material for the corpus be clear, consistent and transparent. Indeed, transparency is key in selecting material for the corpus. The criteria used when selecting material also need to be explicitly stated when reporting to others about a study—it is a basic principle in research and a matter of making it possible for others to replicate the study. The selection criteria are typically biased with respect to specific research interests behind a given corpus project, which should also be spelled out in the documentation about the corpus.

[2] Biber (1993:244) defines a sampling frame as "an operational definition of the population, an itemized listing of population members from which a representative sample can be chosen".

1.2.2 *Issues in Collecting Data for the Corpus*

Corpus compilation involves a series of practical considerations having to do with the question ‘Given the relative ease of access, how much data is it feasible to collect for the corpus?’. Indeed, this needs addressing before it is possible to determine fully the design of a corpus. Relevant spoken or written material may of course be found in many different places, and the effort required to collect it may vary considerably. Some types of discourse are meant to be widely distributed, and are even in the public domain, while others are relatively hidden, and are even confidential or secret. In an academic setting, for example, written course descriptions and spoken lectures target a large audience, while teacher feedback and committee discussions about the ranking of applicants for a post target a restricted audience.

Once the data have been collected, varying degrees of data management will be required depending on the nature and form of the data. If spoken material is to be included in the corpus, it needs to be transcribed, that is, rendered in written form to be searchable by computer. The transcription needs to be appropriately detailed for the research question (see Chap. 11 for key issues involved in compiling spoken corpora). If written material is to be included in the corpus, there are practical considerations regarding how it is encoded. For example, if it can be accessed as plain text files at the time of collection, it will save time. If it is only available on paper, it will need to be scanned using OCR (Optical Character Recognition) in order for the text to be retrieved. If it is only available on parchment, it will need very careful handling indeed by the historical corpus compiler, including manual typing and annotation to represent it. Even modern text files which are available in pdf format may not be retrievable as plain text at all, or it may be possible to convert the pdf to text, but only with a varying degree of added symbols and garbled text, requiring additional ‘cleaning’.[3] Section 1.2.5 on Formatting the corpus discusses some of these issues more fully.

Nowadays there are massive amounts of material on the web, which are already in an electronic format. As a consequence, it has become popular among corpus builders to include material from online sources (see Chap. 15), which represent a great variety of genres, ranging from research articles to blogs. It is important, however, to make the relevance of the material to the research question a priority over ease of access, and carefully consider questions such as “How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared to the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence?” (Sinclair 2005).

Even if material is available on the web, it does not necessarily mean that it is easy to access—at least not in the way texts need to be accessed for corpus work. Online newspapers are a case in point. While they often make it possible to search

[3] There are tools that automatically convert pdf files to simple text, such as AntFileConverter http://www.laurenceanthony.net/software/antfileconverter/. Accessed 24 May 2019.

the archive, they may not make the text files downloadable other than one by one by clicking a hyperlink. The work of clicking the link, copying and saving each individual article manually is then left to the user. This is no small task, but it tends to be underestimated by beginner corpus compilers. Fortunately, there are ways of speeding up and automatizing the process in order to avoid too much manual work; Chap. 15 offers suggestions.

Corpus compilers who are able to collect relevant material in the public domain still need to check the accuracy and adequacy of the material. Consider the case of a research group seeking the answer to the question 'To what extent is (a) the spoken dialogue in the fictional television series *X* (dis)similar to (b) authentic non-scripted conversation?'. They may go to the series' website to search for material, following the logic that an official website is likely to be a more credible source for transcripts than a site created by anonymous fans. Before any material can be included in the corpus, however, each transcript needs to be checked against the recorded episode to ensure that the transcription is not only correct, but also sufficiently detailed for the specific research purposes. When collecting material from the web, there may also be copyright restrictions to take into account; see e.g. the section on Ethical considerations below and Section 3.2 in McEnery and Hardie (2012) on legal issues in collecting such data.

Beginner corpus researchers often find themselves confounded by the question 'How much data do I need in order for my study to be valid?'. There is no rule of thumb for corpus size, except for the general principle 'the more, the better'. That said, it requires more data to be able to make valid observations about a large group of people and a general type of discourse than a small group of people and a specific type of discourse. It also requires more data to investigate rare rather than common linguistic features. Thus, the appropriate amount of data depends on the aim of the research. Each study, however, needs to be considered in its context. There are always going to be practical restrictions on how much time a given researcher is able to put into a project. Researchers who find themselves in a situation of not being able to collect as much data as planned will need to adjust their research questions accordingly. With less data—a smaller sample—the claims one is able to make based on one's corpus findings will be more modest. But most importantly, as discussed above, the issue of representativeness needs to be addressed before a corpus, regardless of size, can be considered appropriate for a given study.

1.2.3 Ethical Considerations

Corpus compilation involves different types of ethical considerations depending on the type of data. For data in the public domain, such as published fiction or online newspaper text, it is not necessary to secure consent. However, such data may be protected by copyright. For data that is collected from scratch by the researcher, it is necessary to obtain the informants' informed consent and it may be necessary to ask for institutional approval.

In the case of already published material, permission may be needed from a publisher or some other copyright holder. There are grey areas in copyright law and copyright infringement is looked at in different ways in different parts of the world, so it is difficult to find universally valid advice on the topic, but generally speaking copyright may prove quite a hindrance for corpus compilation. To a certain extent, restrictions on copyright may be alleviated through concepts such as 'fair use', as texts in a corpus are typically used for research or teaching purposes only, with no bearing on the market.[4] However, copyright holders and judges are likely to distinguish between material that is used by a single researcher only and material that is distributed to other researchers, so it may matter whether or not the corpus is made available to the wider research community. In addition to the potential difference between data gathering for a single use versus data distribution for repeated use by many different people, copyright holders may be more likely to grant permission to use an extract rather than a complete text.

In the case of collecting data from informants, approval may be needed from an institutional ethics review board before the project can begin. Even if institutional approval is not needed, consent needs to be sought from the informants in order to collect and use the data for research purposes. Asking for permission to use material for a corpus is often done by means of a consent form, which is signed by each informant, or by the legal guardians in the case of children (see Chap. 14). A consent form should clearly state what the data will be used for so that an informed decision can be made. It needs to be clear that the decision to give consent is completely voluntary. It is important how the consent form is worded, so it is useful to consider forms used in similar corpus projects for comparison.[5] If a participant does not give his or her consent, the data will have to be removed from the corpus. In the case of multi-party interactions, it may still be worth including the data if most participants have given their consent, while blanking out contributions from the non-consent-giving participant. See Crasborn (2010) for a problematized view of consent in connection with online publication of data.

Once permission has been obtained to use data for a corpus, the informants' integrity needs to be protected in different ways, such as by anonymizing the material. An initial step may be to not reveal the identity of the informants by not showing their real names, for example through 'pseudonymisation', whereby personal data is transformed in such a way that it cannot be attributed to a specific informant without the use of additional information, which is kept separately. A second step may be to manipulate the actual linguistic data (that is, what the

[4] Fair use is measured through the purpose and character of the use, the nature of the copyrighted work, the amount and substantiality of the portion taken, and the effect of the use on the potential market. For more information, see https://fairuse.stanford.edu/overview/fair-use/four-factors/. Accessed 24 May 2019.

[5] For sample templates, see the forms from the *Bavarian Archive for Speech Signals* at http://www.phonetik.uni-muenchen.de/Bas/BasTemplateInformedConsent_en.pdf, or Newcastle University at https://www.ncl.ac.uk/media/wwwnclacuk/research/files/Example%20Consent%20Form.pdf. Accessed 29 May 2019.

people represented in the corpus said or wrote) by changing also names and places mentioned which could in some way give away the source. In the case of image data, this would involve masking participants' identity in various ways.

Confidential data needs to be stored in a safe way. Sensitive information may have to be destroyed if there is a risk that others may access information which informants have been promised will not be revealed. For further reading on ethical perspectives on data collection, see e.g. BAAL's *Recommendations on Good Practice in Applied Linguistics*.[6] It complicates matters that regulations may differ by region. While ethics review boards have been in place for quite some time at universities in the United States, linguists in Europe have been relatively free to collect data. It is not clear, however, what long-term effects the General Data Protection Regulation (https://www.eugdpr.org/; effective as of 2018) will have for data collected in the European Union.

1.2.4 Documenting What Is in the Corpus

As language use is characterized by variability, factors which may have an impact on the way in which language is used should be recorded in some way—these may include demographic information about the speakers/writers, or situational information such as the purpose of the communication or the type of relationship between the discourse participants. Even if the corpus compilers are deeply familiar with the material, it is still the case that memory is both short and fallible, so if they want to use the corpus in a few years' time, important details of the specific context of the data may well have been forgotten. In addition, if the corpus is made available to others, they need to know what is in it in order to make an informed decision about whether the design of the corpus is appropriate for answering their specific research questions.

Anybody who wants the claims made based on a corpus to be accepted by the research community needs to show in some way that the corpus material is appropriate for the type of research done. With incomplete description of the corpus, people will be left wondering whether the material was in fact valid for the study. There are several different ways in which information about the corpus design can be disseminated. It can be done through a research publication, such as a research article or an MA thesis, which includes a section or chapter describing the material (for more on this, see Chap. 26). Corpus descriptions are sometimes published in peer-reviewed journals, especially if the corpus is breaking new ground (as is the case in Representative Study 2 below), so that the research community can benefit from discussions on corpus design. It can also be done by writing a report solely dedicated to describing the corpus (and possibly how to use it), which is made

[6] See https://baalweb.files.wordpress.com/2016/10/goodpractice_full_2016.pdf. Accessed 24 May 2019.

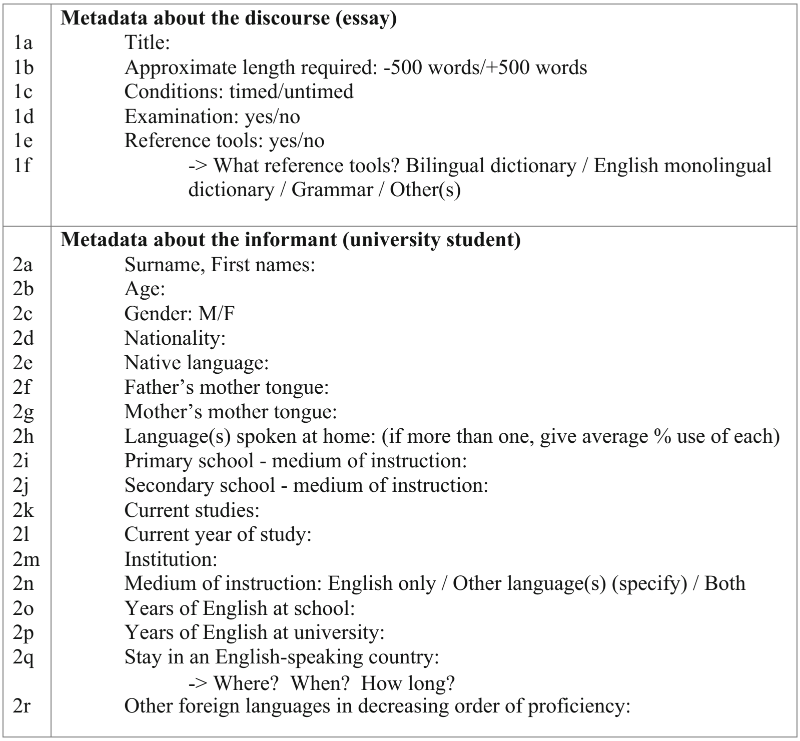
available either as a separate file stored together with the corpus itself, or online. Corpora often come with "read me" files where the corpus design is accounted for. Some large corpus projects intended to attract large numbers of users, such as the *British National Corpus* (BNC) and the *Michigan Corpus of Academic Spoken English* (MICASE), provide relatively detailed reports online.[7] There are also published books which offer even more detailed documentation of corpora and recommendations for how to use them (e.g. Aston and Burnard's (1998) *The BNC Handbook* and Simpson-Vlach and Leicher's (2006) *The MICASE Handbook*).

Another reason for documenting what is in the corpus is to enable researchers to draw on various variables in a systematic way when analyzing data from the corpus. As an example, see Chap. 8 and the subsection on quantitative analysis of concordance lines. In a study of *that*-complementation in English, for each hit in the corpus, the researchers considered external variables such as the L1 of the speaker who had produced the hit and whether the hit came from a written or spoken mode.

Through the inclusion of 'metadata'—data *about* the data—about the type of discourse represented in the corpus, the corpus user can keep track of or investigate different factors that may influence language use, which may explain differences observed in different types of data. Metadata can consist of different types of information. For example, the corpus compiler may include information based on interviews with participants or participant observation. A common way of collecting metadata is by asking corpus participants to fill out a questionnaire which has been carefully designed by the corpus compiler so as to include information likely to be relevant with respect to the specific context of the discourse included and the people represented. An example of metadata based on a questionnaire from the *International Corpus of Learner English* (ICLE) is summarized in Fig. 1.1. The ICLE is a large-scale project with collaborators from several different countries. (For more information on learner corpora, see Chap. 13.) The corpus includes metadata about the type of discourse included (written essays) and about the language users represented (university students), collected through a questionnaire called a 'learner profile', as the contributors are all learners of English. In a language-learning context, some of the variables likely to be relevant include what the learner's first language is (2e), what the medium of instruction was in school (2i; 2j), how much exposure the learner has had to the second language—whether through instruction in a school context (2l) or through spending time in a context where the second language is spoken (2q).

Based on metadata from the questionnaire, it is possible to select a subset of the ICLE corpus, for example to study systematically potential differences in language use between learners who have and who have not spent any time abroad in a country where the target language is spoken natively—and thus test a hypothesis from Second Language Acquisition research.

[7] See http://www.natcorp.ox.ac.uk/docs/URG/ (Accessed 24 May 2019) and https://web.archive.org/web/20130302203713/http://micase.elicorpora.info/files/0000/0015/MICASE_MANUAL.pdf (Accessed 24 May 2019).



	Metadata about the discourse (essay)
1a	Title:
1b	Approximate length required: -500 words/+500 words
1c	Conditions: timed/untimed
1d	Examination: yes/no
1e	Reference tools: yes/no
1f	-> What reference tools? Bilingual dictionary / English monolingual dictionary / Grammar / Other(s)
	Metadata about the informant (university student)
2a	Surname, First names:
2b	Age:
2c	Gender: M/F
2d	Nationality:
2e	Native language:
2f	Father's mother tongue:
2g	Mother's mother tongue:
2h	Language(s) spoken at home: (if more than one, give average % use of each)
2i	Primary school - medium of instruction:
2j	Secondary school - medium of instruction:
2k	Current studies:
2l	Current year of study:
2m	Institution:
2n	Medium of instruction: English only / Other language(s) (specify) / Both
2o	Years of English at school:
2p	Years of English at university:
2q	Stay in an English-speaking country: -> Where? When? How long?
2r	Other foreign languages in decreasing order of proficiency:

Fig. 1.1 An example of metadata collected for a corpus: The learner profile for the ICLE. (Adapted from https://uclouvain.be/en/research-institutes/ilc/cecl/corpus-collection-guidelines.html. Accessed 24 May 2019)

Three different documents that are commonly used in corpus compilation have been brought up above: (i) the consent form from the participants, (ii) the questionnaire asking for various types of metadata about the participants and the discourse and (iii) a text, possibly in a "read me" file, which documents what is in the corpus. Corpus compilers who are collecting publicly available data in such a way that they do not need (i) or (ii), may still choose to compile metadata to help track for instance various types of sociolinguistic information about the corpus participants. However, if both (i) and (ii) are needed, it is a good idea to investigate the possibility of setting them up electronically, such as on a website, to avoid having to type in all the responses manually.

1.2.5 Formatting and Enriching the Corpus

There is a great deal to be said about how best to format corpus material, but this section will merely offer a few hints on technicalities. (More detailed information is found in Chaps. 2 and 3.) Researchers' computational needs and programming skills vary. Those who are reasonably computer literate and whose corpus needs are relatively simple are likely to be able to do all the formatting themselves. However, those who wish to compile a corpus involving complex types of information or do advanced types of (semi-)automatic corpus searches would be helped by collaborating with a computational linguist or computer programmer (see Chap. 9).

A plain text format (such as *.txt*) is often used for corpus files. MS Word formats are avoided, as these add various types of information to the file and do not work with corpus tools such as concordance programs. When naming files for the corpus, it is useful to have the file name in some way reflect what is in the file. For example, the file name 'BIO.G0.02.3' in a corpus of university student writing across disciplines and student levels (*Michigan Corpus of Upper-level Student Papers, MICUSP*; see Römer and O'Donnell 2011), consists of an initial discipline code ('Biology'), a student level code ('G0' stands for final year undergraduate; while 'G1' stands for first year graduate, etc.), followed by a student and paper number ('02.3' refers to the third paper submitted by the second student at that level). Codes not only make it easier for the analyst to select the relevant files, but are also useful when analyzing concordance results, as the codes may help reveal patterns in the data. For example, in studying adverbial usage in student writing, the analyst may find that all of the hits for the relatively informal adverbial *maybe* come from texts coded with the lowest student level ('G0').

It may be necessary, or just a good investment of time, to add markup, that is, "codes inserted into a corpus file to indicate features of the original text rather than the actual words of the text. In a spoken text, markup might include utterance breaks, speaker identification codes, and so on; in a written text, it might include paragraph breaks, indications of omitted pictures and other aspects of layout" (McEnery and Hardie 2012:246). If we take an example from the corpus of university student writing mentioned above, one of the marked-up features is quoted material. This makes it possible to exclude quotations when searching the running text, based on the logic that most corpus users would be primarily interested in text produced by novice academics themselves, and not material brought in from primary or secondary sources.

Markup allows the corpus builder to include important information about each file in the corpus. Various types of metadata can be placed in a separate file or in a 'header', so that a computer script or web-based tool for example will be able to use the information in systematic ways when counting frequencies, searching for or displaying relevant data. If we consider the metadata from the ICLE (Fig. 1.1 above) again, it makes it possible to distinguish for instance between those essays which were timed versus untimed, or between essays written by learners who have never

```
3a  <s n="652">
3b  <w c5="CJC" hw="but" pos="CONJ">But </w>
3c  <w c5="AV0" hw="like" pos="ADV">like</w>
3d  <c c5="PUN">, </c>
3e  <w c5="PNP" hw="i" pos="PRON">I</w>
3f  <w c5="VBD" hw="be" pos="VERB">was</w>
3g  <seg Quotatives="QTG">
3h  <w c5="VVG" hw="think" pos="VERB">thinking</w>
3i  </seg>
3j  <pause/>
3k  <seg Reporting_modes="MDD">
3l  <w c5="DT0" hw="this" pos="ADJ">this </w>
3m  <w c5="VBZ" hw="be" pos="VERB">is </w>
3n  <w c5="VVG" hw="gon" pos="VERB">gon</w>
3o  <w c5="TO0" hw="na" pos="PREP">na </w>
3p  <w c5="VBI" hw="be" pos="VERB">be </w>
3q  <w c5="AV0" hw="so" pos="ADV">so </w>
3r  <w c5="AJ0" hw="embarrassing" pos="ADJ">embarrassing </w>
3s  <w c5="AV0" hw="like" pos="ADV">like </w>
3t  <w c5="PRP" hw="in" pos="PREP">in </w>
3u  <w c5="ZZ0" hw="p" pos="SUBST">P </w>
3v  <w c5="ZZ0" hw="e" pos="SUBST">E</w>
3x  <c c5="PUN">!</c>
3y  </seg>
3z  </s>
```

Fig. 1.2 An illustration of XML annotation: Sentence (a) from the corpus in Representative Study 2. (Based on Rühlemann and O'Donnell 2012:337)

stayed in a country where the target language is spoken versus learners who have reported on relatively extensive stays in such a context.

Another way of adding useful information to a corpus is through annotation, or "codes within a corpus that embody one or more linguistic analyses of the language in the corpus" (McEnery and Hardie 2012:238). Annotation can be done manually or (semi-)automatically (see Chap. 2 for information about automatic annotation). Annotation helps to make the data more interesting and useful. It can be done at any linguistic level, including for example classification of word class for each word in the corpus (POS-tagging; see Fig. 1.2), indication of prosodic features of spoken data, or pragmatic marking of politeness phenomena. Representative study 2 presents annotations of narratives in conversation, which for example involved adding a code for the degree to which an utterance is represented as verbatim or indirect. Example (a) from the corpus includes a sentence from an utterance where the underlined unit is coded 'MDD' (3k in Fig. 1.2) for a verbatim presentation mode.

(a) But like, I was thinking this is gonna be so embarrassing like in P E!

The contemporary standard for corpus markup and annotation is XML (eXtensible Markup Language), where added information is indicated by angle brackets <>, as illustrated in Fig. 1.2, which represents the above sentence. The sentence opens with an <s> tag including a number which uniquely identifies it (3a), and closes with an end tag </s> (3z). Each word also has an opening <w> tag, giving information about lemma forms and part of speech ('pos'), and a closing </w> tag. The quotative verb (3h), for example, is labelled "VERB" and, more specifically, "VVG" to mark the *–ing* form of a lexical verb. We can also see, for example, that 3f (*was*), 3m (*is*) and 3p (*be*) instantiate different forms of the lemma BE.

XML is ideal "because of its standard nature" and "because so much corpus software is (at least partially) XML-aware" (Hardie 2014:77–78). This does not mean, however, that it is necessary to use in corpus building. While Representative Study 2 is at the advanced end regarding corpus formatting, Representative Study 1 uses raw corpus texts and does not even mention XML or annotation. The degree to which a corpus is enriched will depend partly on the research objectives. MICUSP was mentioned above as an example of a corpus created with the aim of mapping advanced student writing across different levels and disciplines. As mentioned, quoted material was marked up to enable automatic separation between the students' own writing and writing from other sources. It is also an example of a corpus that is distributed to others, which means that the compilers put a greater effort into marking up the data for a range of potential future research projects. For those wishing to learn more about XML for corpus construction, Hardie (2014:73) is a good place to start.

Even more fundamental than markup or annotation is encoding, which refers to "the process of representing a text as a sequence of characters in computer memory" (McEnery and Hardie 2012:243). We want corpus texts to be rendered and recognized the same way regardless of computer platform or software, but for example accented characters in Western European languages (such as *ç* and *ä*) may cause problems if standard encoding formats are not used. How characters are encoded may be an issue especially for non-alphabetical languages. A useful source on the fundamentals of character encoding in corpus creation is McEnery and Xiao (2005), who recommend the format UTF-8 for corpus construction, as it represents "a universal format for data exchange" in the Unicode standard. Unicode is "a large character set covering most of the world's writing systems, offering a way of standardizing the hundreds of different encoding systems for rendering electronic text in different languages, which were often conflicting" (Baker et al. 2006:163) in the past. Unicode and XML together currently form a standard in corpus building.

There are many considerations for formatting corpus material in ways that follow current standards and best practice. An authoritative source is the Text Encoding Initiative (TEI),[8] which represents a collective enterprise for developing and maintaining international guidelines. The TEI provides recommendations for different aspects of corpus building, ranging from how to transcribe spoken data to what to put in the 'header'. As mentioned above, some corpus projects make use of 'headers' placed at the top of each corpus file. A TEI-conformant header

[8] See https://tei-c.org/. Accessed 24 May 2019.

should at least document the corpus file with respect to the text itself, its source, its encoding, and its (possible) revisions. This type of information can be used directly by linguists searching the corpus texts, but most often it is processed automatically by corpus tools to help the linguist pre-select files, visualize the distribution of variables, display characters correctly, and so on.

1.2.6 Sharing the Corpus

One of many ways in which corpora vary is in how extensively and long-term they are intended to be used. A corpus can be designed to be the key material for many different research projects for a long time to come, or it can be created with a single project in mind, with no concrete plan to make it available to others. In the former category, we find 'standard' corpora, which are widely distributed and which form the basis for a large body of research. This type of corpus is designed to be representative of a large group of speakers, typically adopting "the ambitious goal of representing a complete language", as Biber (1993:244) puts it. In the latter category, we find a large and ever-growing number of corpora created on a much more modest scope, focusing on a small subset of language. These are oftentimes used by a single researcher to answer one specific set of research questions, as in the case of Representative Study 1.

Even in the context of a small-scale corpus project, it is considered good practice in research to make one's data available to others. It supports the principle of replicability in research and it fosters generosity in the research community. Our time will be much better invested if more than one person actually uses the material we have put together so meticulously. Certain types of data will be of great interest to not only researchers or teachers and students, but also the producer community itself, as in the case of sign language corpora (e.g. Crasborn 2010). Sharing one's corpus is in fact to an increasing extent a requirement; some bodies of research funding make 'open access' a precondition for receiving any funding. When sharing a corpus, it is common to apply licensing. Making a corpus subject to a user licence agreement provides a way of keeping a record of the users and of enforcing specific terms of use. Corpora published online may for example be made available to others through a Creative Commons licence in order to prohibit profit-making from the material.[9] However, even with such a licence in place, it may be difficult for corpus compilers to enforce compliance, which is another reason for taking very seriously the protection of informants' integrity.

Even if open access is not a requirement, in a case where a researcher is applying for funding to compile a corpus for a research project, it may be a good idea to include an entry in the budget for eventually making the corpus available. If, say for various reasons related to copyright, it is not possible to make the complete set of

[9] See https://creativecommons.org/licenses/. Accessed 24 May 2019.

corpus files available to others, the corpus could still be made searchable online and concordance lines from the corpus be shown.

Another consideration in sharing corpus resources involves how to make these accessible to others and how to preserve digital data. The easiest option is to find an archive for the corpus, such as The Oxford Text Archive or CLARIN.[10]

1.2.7 Corpus Comparison

Corpus data are typically studied quantitatively in some capacity. This means that the researcher will have various numbers to which to relate, which typically give rise to questions such as ‘Is a frequency of *X* a lot or a little?’. Such questions are difficult to answer in a vacuum, but are more usefully explored by means of comparison—for example by studying the target linguistic phenomenon not just in one context, but contrasting it across different contexts. Statistics can then be used to support the interpretation of results across two or more corpora, or to assess the similarity between two or more corpora (see e.g. Kilgarriff (2001) for a classic paper taking a statistical approach to measuring corpus similarity).

The researcher may go on to ask qualitative questions such as ‘How is phenomenon *X* used?’ and systematically study similarities and differences in (sub-)corpus A and (sub-)corpus B. Even if frequencies are similar in cross-corpus comparison, it may be the case that, once you scratch the surface and do a qualitative analysis of how the individual examples are actually used, considerable differences emerge. In order for the comparison to be valid, however, the two sets ((sub-)corpus A and (sub-)corpus B) need to be maximally comparable with regard to all or most factors, except for the one being contrasted.

Some corpora are intentionally constructed for comparative studies (this includes parallel corpora, covered in Chap. 12). In contrastive studies of different languages or varieties, for example, it is useful to have a so-called comparable corpus, which “contains two or more sections sampled from different languages or varieties of the same language in such a way as to ensure comparability” (McEnery and Hardie 2012:240). The way in which the texts included in the corpora have been chosen should be identical or similar—that is, covering the same type of discourse, taken from the same period of time, etc.—to avoid comparing apples to oranges.

Having considered some of the fundamentals of corpus compilation, we will next turn to the two sample studies, which will illustrate further many of the concepts mentioned in this section.

[10] See https://ota.ox.ac.uk/ (Accessed 24 May 2019) and https://www.clarin-d.net/en/corpora (Accessed 24 May 2019).

Representative Study 1

Jaworska, S. 2016. A comparison of Western and local descriptions of hosts in promotional tourism discourse. *Corpora* 11(1): 83–111.

Jaworska (2016:84) makes the point that "corpus tools and methods [are] increasingly used to study discursive constructions of social groups, especially the social Other—that is, groups that have been marginalised and discriminated against".[11] In this study, corpus methods are used to investigate promotional tourism discourse and ways in which local people (hosts) are represented. Previous research in the area is based on small samples of texts and looks at representations in one destination or region, so there is typically no comparison across contexts. The research questions for the study are:

1. How are hosts represented in tourism promotional materials produced by Western versus local tourist industries?
2. To what extent do these representations differ?
3. What is the nature of the relationship between the representations found in the data and existing stereotypical, colonial, and often gendered ideologies?

To answer these questions, two corpora were created, consisting of written texts promoting tourist destinations that have a history of being colonised. The two corpora represent, on the one hand, a Western, 'external' perspective and, on the other, a local, 'internal' perspective, which are contrasted in the study. They are labelled the *External Corpus* (EC) and the *Internal Corpus* (IC).

To create the EC, texts were manually taken from the websites of "some of the largest tourism corporations operating in Western Europe". A selection of 16 destinations was made, based on the most popular destinations as identified by the companies themselves during the period of data collection—however excluding Southern European destinations, as the focus of the research was on post-colonial discursive practices. To create the IC, official tourism websites were sourced from the 16 countries selected in the process of creating the EC. All of the websites are listed in an appendix to the article.

A restriction imposed on the data selection for both corpora was to include only "texts that describe the countries and its main destinations (regions and towns)" rather than specific resorts or hotels or information on how to get there. This was to make the two corpora as comparable as possible. However, one way in which they differ is with respect to size, with the IC being three

(continued)

[11] A similar example is reported in Chap. 8 and involves a study of how foreign doctors are represented in a corpus of British press articles.

times as big as the EC, as "local tourism boards [offer] longer descriptions and more details" (92). The solution to comparing corpora of different sizes was to normalise the numbers, rather than reduce the size of the IC. The author's rationale was that reducing the IC would have "compromise[d] the context and the discourse of local tourism boards in that some valuable textual data could have been lost" (92).

The corpora were compared by extracting lists of the most frequent nouns (cf. Chap. 4). From these lists were identified the most frequent items used to refer to local people (e.g. *people*, *locals*, *man/men*, *woman/women*, *fishermen*). Careful manual analysis was required in order to check that each instance was relevant, that is, actually referring to hosts/local people. The word *people*, for example, was also sometimes used to refer to tourists. It was found that the IC had not only more tokens of such references, but also more types (F = 68) compared to the EC (F = 20). The tokens were further classified into socio-semantic groups of social actors based on an adapted taxonomy from the literature, for example based on 'occupation' (*fisherman*, *butler*), 'provenance' (*locals*, *inhabitants*), 'relationship' (*tribe*, *citizens*), 'religion' (*devotees*, *pilgrims*), 'kinship' (*son/s*, *child/ren*) and 'gender' (*man/men*, *woman/women*).

The corpora were compared qualitatively as well, by identifying patterns in the concordance lines and analysing the context ("collocational profiles") of the references to hosts, specifically of *people* and *locals*, which occurred in both corpora. The pattern found for *locals* was that local people were represented "on an equal footing with tourists" in the IC, while in the EC they were portrayed as "docile, friendly and smiley servants [,] reproduc[ing] and maintain[ing] the ideological colonial asymmetry" (104).

Representative Study 2

Rühlemann, C. and O'Donnell, M.B. 2012. The creation and annotation of a corpus of conversational narratives. *Corpus Linguistics and Linguistic Theory* 8(2): 313–350.

Rühlemann and O'Donnell's (2012) article *Introducing a corpus of conversational stories: Construction and annotation of the Narrative Corpus* describes the main features of a corpus of conversational narratives. Research has shown that it is extremely common for people to tell stories in everyday conversation. The authors hope that the use of the corpus "will advance the linguistic theory of narrative as a primary mode of everyday spoken interaction" (315).

(continued)

Previous work on this type of discourse has been based not on corpus data, but on elicited interviews or narratives told by professional narrators.

The corpus comprises selected extracts of narratives, 153 in all, for a total of around 150,000 words, taken from the demographically sampled ‘casual conversations’ section of the BNC, which is balanced by sex, age group, region and social class, and which totals approximately 4.5 million words. This example is somewhat unusual in that the authors do not collect the data themselves, but instead use a selection of data from an existing corpus. However, given that the intended audience of this handbook is expected to have limited resources for corpus compilation, it seems useful to provide an example of a study where it was possible to use part of an already existing corpus. The NC is only about 3% of the original collection from BNC, so the authors have put a great deal of effort into selecting the data, which is done in a transparent and principled way. In the article, they describe (i) the extraction techniques, (ii) selection criteria and (iii) sampling methods used in constructing the corpus. In order to (i) retrieve narratives, they (a) read the files manually and (b) used a small set of lexical forms (e.g. *it was so funny/weird*; *did I tell you*; *reminds me*) that tend to occur in narratives based on the literature or based on analysis of their own data. In (ii) deciding what counts as a conversational narrative, they used three selection criteria: First, some kind of ‘exosituational orientation’ needed to be present in the discourse, that is, “linguistic evidence of the fact that stories relate sequences of events that happened in a situation remote from the present, story-telling, situation” (317)—this includes for example the use of past tense verbs; items with past time reference as in *yesterday*; reference to locations not identical to the location of speaking. A second criterion was that at least two narrative clauses be present, which are temporally related so that first one event takes place and then another. A third criterion involved consensus, so that at least two researchers agreed that a given example was in fact a narrative. With respect to (iii) sampling, the authors retained the sociological balance from the demographically sampled BNC by choosing two texts from each file insofar as this was possible.

The NC is not only a carefully selected subset of the demographically sampled BNC, but it is also annotated. The corpus builders have thus augmented the existing data by adding various types of information—about the speakers (sex, age, social class, region of origin, educational background), about the text (type of narrative; whether a stand-alone story or part of a ‘narrative chain’) and about the utterance (the roles of the participants vis-à-vis the narration; type of quotative verb used to signal who said what in a narrative; to what degree the discourse is represented as being verbatim or more or less indirect). The authors stress that all of the annotation is justified in some way by the literature on conversational narrative, so the rationale for

(continued)

including a layer of analysis to the corpus text is to enable researchers to answer central research questions in a systematic fashion.

The corpus design makes it possible to use the demographic information about the speakers—such as sex—and consider how it is distributed in relation to the number of words uttered by the speakers who are involved in the narratives, as exemplified in Table 1.1. Note the presence of a category of "unknown", which is useful when relevant metadata is missing.

Each narrative in the corpus is classified also based on a taxonomy of narrative types. This type of information is highly useful, as it not only makes it possible to study and compare different types of narrative, but it also shows how the corpus is balanced (or not) with respect to type of narrative. The classification is justified by an observation from the literature that "we are probably better off [] considering narrative genre as a continuous cline, consisting of many subgenres, each of which may need differential research treatment" (Ervin-Tripp and Küntay 1997:139, cited in Rühlemann and O'Donnell 2012:321). The annotation includes two features: experiencer person (whether first person or third person, that is, direct involvement by narrator versus hearsay) and type of experience (personal experiences; recurrent generalized experiences; dreams; fantasies; jokes; mediated experiences). The last subcategory refers to the common practice of retelling a film or a novel.

At the time of creation, the NC was the first corpus of conversational narratives to be annotated, so there was no established practice to follow regarding what analytical categories to annotate. However, the authors were able to follow some general guidelines, for example Leech's (1997) 'standards' for corpus annotation concerning how to design the labels in the tagsets (e.g. they should be (a) easy to interpret and (b) concise, consisting of no more than three characters).

Table 1.1 Distributions of male and female narrative participants involved in narratives [based on a subset of the total corpus]

Sex	Number of participants	%	Number of words	%
Female	212	42	44,476	56
Male	173	35	24,268	31
Unknown	115	23	10,079	13
Total	**500**	**100**	**78,823**	**100**

Rühlemann and O'Donnell (2012:320)