1 Introduction

1.1 Genesis of the *EDD*-Related Projects

EDD Online is a digitised version of Joseph Wright's *English Dialect Dictionary*, which was published from 1898 (or 1896[1]) to 1905 and covers the time from 1700 to 1904, with occasional, mainly etymologically motivated references to the preceding 1,000 years, that is, from 700 to 1700. The digitisation naturally implies OCR (optical character recognition). The paper version of the dictionary comprises some 4,600 pages, densely printed in two columns. The dialects depicted are those of the British Isles, historically including the whole of Ireland as well as the USA, Canada, Australia, South Africa and 'colonial' Englishes. The dictionary also includes what, from a later point of view, would be called 'sociolects', namely colloquial English, slang and cant, as well as 'technical' Englishes as used by professional groups such as farmers, miners and so on.

EDD Online is also the name of the last two Innsbruck *EDD*-related projects, supported by the *Austrian Science Fund* and aiming to provide an optimally digitised version of Wright's *Dictionary*. The first of the two projects lasted three years (1 April 2011 to 30 March 2014). The follow-up project, called *EDD Online: Applied, Corrected and Supplemented*, was carried out over eighteen months from 2017 to 2018. Both projects were based on the work of a previous project, *SPEED* (*Spoken English in Early Dialects*), which ran from 2006 to 2010 and had the purpose of presenting a preliminary online version of the *Dictionary*. The text of this initial phase was produced by a scanner and was not proofread, with the result of several unavoidable misreadings. The subsequent (automatic) tagging of the text was, naturally, insufficient, as was the interface provided for this data. But this first project was, in hindsight, an absolutely vital 'dress rehearsal' of our later performance.

[1] The earlier date is justified if one counts the pre-published fascicles, i.e. parts of the first volume. Part I – *A* to *Ballot* – was published on 1 July 1896 (cf. E. M. Wright 1932: II, 397). Volume I of the *EDD* in its complete form came out in 1898.

The two projects that have *EDD Online* in their names were, thus, the necessary follow-up projects of *SPEED*, with the aim of correcting the mistakes in *SPEED*, both of orthography and tagging. Moreover, the purpose of the follow-up projects was substantially to improve our interface, i.e. the surface on the screen allowing access to the *Dictionary*, in accordance with our corrections and with quite a number of better insights on what researchers can and will wish to do with Wright's *EDD*.

The official time of the second project was extended twice, each time for a year. The last project was planned for a year and a half from the very beginning. These extensions were unavoidable due to the enormous problems involved and despite our hard and keen work over the years. However, we did not require more money on our way, only more time. To put it in a simile close at hand in a town like Innsbruck, our work was like a mountaineering tour to an unknown peak, where one can never reliably predict when exactly one will reach it and when hard endeavour is finally rewarded by a fantastic view. Indeed, this is what *EDD Online*, after a long span of more than ten years of work on my part, finally offers: a fantastic view.

The changing names of the three *EDD*-related projects at issue here deserve a final remark. The names were mainly conditioned by the requirements of the Austrian Science Fund to which I gratefully owe the financial support. While the initial name, SPEED (for *Spoken English in Early Dialects*) was, in hindsight, an eye-catching misnomer, the last name has been too long to be worth remembering. So the two names may now be forsaken in favour of the names borrowed from the second phase of our project work so that we have the sequence *EDD Online 1.0*, *EDD Online 2.0* and *EDD Online 3.0*. Needless to say, when we now refer to *EDD Online* without adding the version, we, naturally, mean the output of the last of the three projects, *3.0*. A recent paper by Markus (2019a) has described the essential additions and innovations provided by version 3.0.

1.2 Overall Structure of the *EDD*

Apart from its paraphernalia (e.g. introduction, list of abbreviations, bibliography, *English Dialect Grammar*), the six volumes of the *EDD* (without the *Supplement*) that consist of some 64,500 entries are alphabetically arranged. To be more precise, the book comprises 71,484 headwords in 64,486 entries. The exactly 4,505 pages of the original dictionary do not include the *Supplement*, which is presented on pages 1–179 of volume VI because this contains, according to Wright, material 'the authority for which was not sufficient' (vol. VI, p. 1 of Suppl.). It also has a structure somewhat deviant from that of the main part of the *Dictionary*. We have, however, integrated Wright's list of *Corrigenda*, marking the necessary emendations

in the XML-version of the running text by an attribute *sicCorr="true"*. Users finding passages in *EDD Online* that deviate from the original paper version of the *Dictionary* should, therefore, check the *Corrigenda* before they reclaim an error.

The later inclusion of the *Supplement* meant the addition of nearly 8,000 entries, which amount to about 11 per cent of the complete dictionary. The details of what the *Supplement* offers and in what way are described by Markus (2019b).[2] In a nutshell, the *Supplement* has added either new entries or new, in particular, semantic information on entries listed in the first place. Some entries are marked as questionable due to 'unsatisfactory authority', but the larger part of the entries simply provides new material.

All in all, the work of Wright and his team has turned out to be admirably scrupulous, knowledgeable and reliable. Nonetheless, it fairly soon became obvious in our correction work letter-wise that the *Dictionary* is not totally homogeneous in its use of descriptive features. Historical and etymological comments, as well as negative or somewhat half-hearted remarks, such as 'common in many parts of x', are considerably more frequent in letters *A* and *B* than in the other letters of the alphabet. Moreover, the *Dictionary's* complex syntax, i.e. the relationship of parts of entries to each other, turned out to be less consistent in the first few letters of the alphabet than with later ones, which suggests that Wright and his team were initially insecure in how to come to terms with the complexity of the data. Part of the 'learning process' on the part of the lexicographers seems to have been their elimination of an undue amount of phonetic or phonemic data (which we find on the first pages of the letter *A*) and to be more resolute concerning the in- or exclusion of material, whereas in *A*, there is still a striking number of comments and additions ranked to be of secondary relevance, mostly added in parentheses or brackets, with the result that these parts of the *Dictionary* are often structured less stringently than the others. Moreover, Wright, during his work on the *Dictionary*, seems to have changed his lexicographic method: from that of a nineteenth-century 'neo-grammarian' (in the way of, say, Alexander Ellis) towards that of a twentieth-century 'structuralist', with F. de Saussure *ante portas*.

Given all these factors of heterogeneity and inconsistency, we were initially (and occasionally) misguided by the first letters of the alphabet in coining tags which we later found to be rather irrelevant for the rest of the *Dictionary*. Another source of confusion was the inconsistent practice of abbreviating, in particular, concerning sources. A book reference to an author's name (like Bunyan), plus two words of the title, such as *Pilgrim's Progress*, may have been verbalised in up to ten different abbreviative versions. This, however,

[2] For a detailed discussion of the *Supplement*, see Markus (2019b).

came as no surprise, given that the *Dictionary* was compiled or prepared from the early 1890s up to 1905, i.e. for more than ten years, and that, naturally, typewriters (not to mention computers) were not available to handle the mass of data. All the more Wright's intuition in using abbreviative codes is to be admired, for example, in the reference to the main sources by way of indexed three-letter codes, such as 'Yks.13', where the numbers stand for titles 1 and 3 of the glossary books listed for Yorkshire in the bibliographical reference list.

1.3 Organisational

EDD Online (2.0) was started in the spring of 2011 when I had previously retired from my active work as a full professor of English linguistics and pre-1,500 literature at the University of Innsbruck. Given the situation in my department as it was, I decided to manage the new project myself, with practically daily presence in our project room in the University.

The project was funded by the Austrian Science Fund with the amount of €299,000, of which the University subtracted a lump sum of €50,000 for infrastructural costs. My team members and I accordingly obtained a project room and all the necessary equipment including computers and software from the University. We mainly used the XML-browser OXYGEN 14.1 and, later, 14.2, for which we got a licence from our computer centre. For programming we used, among other tools, XQuery, HTML, CSS (Cascading Style Sheets), JavaScript and Netbeans.

The follow-up project *EDD Online: Applied, Corrected and Supplemented* (i.e. *EDD Online 3.0*), running from April 2017 to October 2018, was a minor 'ORD project' of the Austrian Science Fund, the abbreviation *ORD* standing for an 'Open Research Data' pilot program,[3] with a budget of €61,000.

1.4 Survey of This Book

The purpose of this book is threefold. First, it familiarises readers with the diverse tools of *EDD Online* (3.0). Given the complexity of our interface (in line with the substance and structure of Wright's *EDD*), this 'handbook' part of the book by far outreaches the practical hints in the short *Guide* provided as part of the interface itself (cf. Markus 2017b). Second, by its going into depth on details of programming (though always from the linguist's and philologist's point of view), the present book wishes to address IT-specialists and laypeople working in philological projects of computerisation, in particular, in the digitisation of dictionaries. Describing the potential, but – in all frankness –

[3] This aims at latest technical standards in the digital age.

also the drawbacks and problems my team and I at the University of Innsbruck have been confronted with is meant to help other computer linguists solve their own problems if these are of a similar kind. With its occasional harping on technical detail, the book is also a documentation of the major part of our Innsbruck expertise as regards *EDD Online* – only after a year, many of these details will be only vaguely remembered or completely forgotten. Third, the book tries to have some impact on both English (computerised) lexicography and dialectology. While I strongly admire many of the outstanding achievements in the past of these two fields, I am also convinced that the new research tools now generally available can and should motivate us to practise new methods.

The method of this book is generally inductive rather than deductive. We will proceed from practical issues close at hand, such as those resulting from the orthography in the *EDD*, to more general topics concerning the use of *EDD Online*, the methods used for creating our data base and its dialectological and linguistic potential, with the overall line of thought more and more moving towards theoretical questions of dialectology and lexicography.

The book has, apart from the Preface, ten chapters. The first, introductory chapter familiarises the reader with the genesis of the Innsbruck project in its different phases and with the main aspects of the structure of the *EDD* as well as the organisational frame of our work in Innsbruck. It concludes with a survey of this book.

The subsequent three chapters (Chapters 2–4) proceed from the text's smaller units, such as the hyphen, special characters, issues of format and mistakes subject to emendation, to the important issue of tagging the original text in XML (Chapter 3) and the syntax of *EDD* entries in the face of the inherent hierarchy between their parts. The rules of TEI (the *Text Encoding Initiative*) are described as an answer to this hierarchical structure.

Chapter 5 mainly aims at corpus linguists and lexicographers as target groups, making practical suggestions based on our project work over the last fifteen odd years. The chapter provides some insights into the way we have digitised and tagged the text as well as a discussion of the software used for developing query commands. It will end with a flow chart purveying an idea of the division of labour practised in our Innsbruck project.

Chapter 6 throws light on the various functions of the *EDD Online* interface, proceeding from simple searches for headwords to the basic functional buttons and icons that are provided in and around the retrieval window (on the left of the interface screen) and then, in the way of a manual, discussing all the functions around the entry window (on the right of the interface screen). Since the icons for the filters are positioned there, 'around the entry window' means that the eight filters available in *EDD Online* are topicalised here in detail.

Chapter 7 returns to, and goes into depth on, the different sub-menus of the retrieval window (on the left of the interface), now with a focus on the 'advanced mode' parameters. By 'parameters' I here mean the types of text units that users can search for. They range from definitions and citations to compounds or other types of word formation and to phrases. Depending on which of the parameters has been opted for, the rules for the combination with other parameters and with filters vary.

Chapter 8 directs our attention to some exemplary research issues within English dialectology and language history to be encouraged by the search tools of *EDD Online*. The chapter first discusses the essence and *raison d'être* of (English) dialectology and then goes on to investigate, partly tentatively, some test cases offering themselves in the face of *EDD Online* as a new tool. A wide range of topics and tasks of dialectology are reflected in eight short studies, ranging from the 'UFO'-quality of dialectal variants to the ubiquity of Shakespeare in dialectal text and lexis. Given that the study of traditional dialects has never overcome its neo-grammatical nineteenth-century background, with its focus on individual words and the word forms of individual dialects, Chapter 8 may, hopefully, inspire dialectologists to try out new, more meaningful and less eclectic approaches.

One of the neglected aspects of traditional dialectology comes to bear in Chapter 9: quantification. The role of a dialectal form or meaning naturally depends very much on aspects of frequency. The counting of dialectal usage items is important in order for us to avoid unjustified generalisation and yet to come to terms with sub-systems of dialectal language. At the same time, measuring frequency presupposes clarity on what is being measured, and with what yardstick. Given that the *EDD* does not represent all counties and areas of the UK and of the English-speaking world alike and objectively, but with clear favourites, the chapter introduces and discusses the ways of normalising frequency figures, relating them to various sum totals. The visualisation of statistics on *ad-hoc* maps as they are provided in *EDD Online* is a further aspect of counting data with good reason and of interpreting them.

The final chapter (Chapter 10) provides, apart from an outlook, short information on the (general) availability of *EDD Online*. Readers of this book are strongly encouraged to test the interface and, thus, to fathom out the real potential of modern dialectology.

2 Orthography

2.1 General

We have tried to keep orthographic mistakes to a minimum by creating first a machine-scanned version of the whole text (in fact, the one used in the project *SPEED*, *alias EDD Online 1.0*) and then a double-typed version – typed by employees of a firm in China. The three versions were then automatically compared (by Hans-Werner Bartz of the University of Trier, later of the *Akademie der Wissenschaften* in Darmstadt, Germany), with a protocol listing the deviant passages so that we could check these passages and correct the mistakes manually. Generally, the mistakes made by the machine were different in type from the mistakes of the human typists. But talking about their mistakes, one should mention that the Chinese typists seemed to have a specific sense of deciphering subtle differences in graphic signs (which is what Chinese spelling consists of), in my opinion, more than educated European typists would, whose minds would probably have read and typed texts based on some sensible (but possibly incorrect) interpretation of words.[1] The *EDD*, of course, created specific problems of spelling semiotics: phonetic transcriptions (of a kind unknown today), with many special characters; pseudo-phonetic spellings (as were widely, but inconsistently common in nineteenth-century Britain); and, last but not least, problems caused by the wide use of abbreviations and the separation of words in line-, column- and page-breaks.

Unlike the main body of the *Dictionary*, the 179 pages of the *Supplement* were not produced in three versions, but only in a machine-scanned one, then to be manually corrected/proofread. This work filled part of the latest phase of the Innsbruck *EDD*-related project and in 2017, as before, was carried out within the XML-version of the text. The editor of OXYGEN 14 was again an

[1] This may be speculative reasoning, derogatory to Western typists, as one of my peer-group referees critically remarked. However, I had to take a decision on the optimal method of reproducing the text, and decisions are sometimes based on (prejudiced) experience.

THUMMEL-POKE, *sb.* Cum.[14] [pu·mɪ·pwok.] A cloth bandage to protect a sore finger, made like a glove and tied with strings round the wrist. (s.v. Huv(v)el.)

Figure 2.1 Use of the extra-short hyphen for compounds (example: THUMMEL-POKE)

appropriate tool for getting on with this task since many systematic scanning mistakes could be easily corrected globally.

2.2 The Problem of the Short Hyphen

Given that the *EDD* implies documentation of the use of mainly spoken English, the spelling of words was bound to be a problem, even to Wright himself. To give the first example: Wright generally used an extra-short hyphen for marking the elements of lemmatised compounds, as if he was uncertain about the value of these almost dot-like hyphens. See the headword in Figure 2.1.

Since dialect lexemes have always been mainly spoken words, Wright could not resolutely decide whether a compound such as *thummel(-)poke* was, or should be, hyphenated or not. The inconsistent use of the hyphen in the text citations, often added to illustrate the lemmas, reveals why Wright seemed unwilling to commit himself in the matter: separate spellings change with joined and clearly hyphenated spellings. In the face of this, Wright's extra-short hyphens in compounds occasionally seem intended as mere markers of morphemic boundaries. Unfortunately, there is no passage in the *EDD*'s editorial comments that explains this practice in further detail.

For a while we considered the possibility of keeping the extra-short hyphens, but then we anticipated the users' problems of interpreting them and, in queries, of typing the 'dubious dot' as a special character. We finally decided for just one type of hyphen, the normal one used today, in addition, of course, to the (somewhat longer) dash, which regularly stands for the substitution of a full word. The sign for the standard hyphen, thus, signals both a morpheme boundary and a possible hyphen proper.

2.3 Special Characters

Special characters were another general problem in reproducing the *EDD* on the computer. Using Unicode characters and, moreover, the coding system of TUSTEP (created at the University of Tübingen by Kuno Schälkle and Wilhelm Ott), we generated almost all the characters and diacritics we needed.

Only very rarely was a separation of diacritics from their graphs unavoidable, for example, initially for some short or long vowels, with both a hook and a stroke above them. But this problem was later solved by our programmers. We did, however, correct Wright's general merger of the *a*- and *o*-ligatures when italicised (<æ> misleadingly looking like <œ>), though this is a general problem that even present-day Microsoft WORD has left unsolved. We kept the two phonemes apart as best we could, based on phonological and etymological reasoning in each single case.

Another special problem we had to tackle was the merger of apostrophes and quotation marks in our initial scanning process. Wright has the two curbed single quotation marks for quotes (example: 'text'), sometimes with spaces before and after the text (' text '), and an equally curbed stroke for the apostrophe (’), which, on the pages of the book, often looks rather like a straight stroke ('). As could be expected, these signs were hopelessly mixed up. To have a clear distinction between the quotation marks and the apostrophe, we regularised the apostrophe to the single stroke ('), the more so since this sign is on the keyboard and does not need to be provided from a list of special characters.

In various other cases, special characters could not be avoided. Thus the ampersand sign '&' had to be coded (&), just like the protected space between parts of headwords ( ) and various phonemic or phonetic symbols with ligatures, accent or stroke superscripts, to mention only these few cases.

2.4 Format

In the *EDD* entries, there are extra-wide spaces (so-called 'spatia') for keeping certain sections of the entries apart. The difference to normal spaces is, however, so minimal that both our scanner and the Chinese typists failed to reproduce them. We did not see sufficient reason for manually reproducing this specificity of format, unlike various other format items (such as boldface), for two reasons (apart from the difficulty of tracing the spatia). (a) In our XML-version of the text (XML = 'Extended Markup Language') each single word has all the functional attributes ('tags') it needs for the researcher's retrieval – see Chapters 3 and 4. All the *prima-facie* formatting of the *EDD* in book form, including font size, types of fonts, small capitals and so on, are, therefore, redundant text features and were seen by us as secondary. (b) The user of *EDD Online* can always switch to the original image of an entry to check what a specific passage or word in the entry looks like. This option of the possible investigation of the original meant a large amount of extra work for us: each word in each line had to be given its specific coordinates in terms of the four corner points of the image. But we have always wished our output to be subject

to users' immediate control. With all this particular care taken by us to allow reproducing the original entry and, moreover, a special string in an entry, we invite all users to inform us about mistakes, should any still be found, or inconsistencies in our text in relation to the original image.

2.5 Emendations

Only very rarely did we have to correct what we found in black and white in the *Dictionary*. Emendation proper has been applied in cases of clear misprints, for example in the case of *brist e-fern*, where an obviously missing <l> (*bristle-fern*) has not come out in the printing (in the entry MAIDEN). We did not comment on such cases, which can be easily interpreted in their contexts, but silently emended these rare passages, not without leaving traces in the XML-version of the text by adding the attribute *sicCorr="true"*. Of course, attributes are not visible to the users in our interface, but can only be checked in the XML-text, which we will keep under strict custody.[2]

Moreover, where there have been inconsistencies in the entries, for example, in the counting of items (e.g. when there was a (2), but no (1)), we have likewise emended the mistakes because they would have irritated the computer's query routines. Such corrections have also always been marked by the attribute *sicCorr="true"*. The same holds true for (rare) emendations caused by 'retrieval deficits'. For example, when the *EDD* referred to a numbered compound as part of, and only used within, a phrase, we had to mark both the compound and the phrase with the given *numerus currens* for the computer to find the quotations and sources attributed to the compound in the subsequent paragraph of the text.[3] The number of the compound was, therefore, repeated by us before the phrase and marked, as all insertions, by double curly brackets to signal the emendation. The interface screen, by the way, presents such emendations by their grey, rather than black, colour.

In addition to these emendations, we had to think of the mistakes corrected by Wright himself. He added half a page of *Corrigenda* at the end of his *Dictionary*, after the *Supplement* (vol. 6, page 179). We have integrated all

[2] This is only meant to exclude non-transparent competitive modifications of our interface. Further enhancement of the work done in Innsbruck in the form of joint ventures would, of course, be welcome.

[3] For example, a compound x listed under (14) in an entry is only attested as part of phrase xy. The quotation and/or source added in the subsequent paragraph under (14) would only be found by the computer in relation to compound (14), but not in relation to the phrase as part of which the compound has survived. The quotation and its source, however, apply to the phrase as much as to the compound.