Christina Sanchez-Stockhammer

Copy and Write: The Transformative Power of Copying in Language

Abstract This contribution explores what can be considered an original and what can be defined as a copy in language. To this end, it elaborates on the role of the classical Saussurean dichotomy langue/parole and factors such as the size and frequency of prefabricated chunks. Furthermore, it discusses how similar linguistic copies can be to a supposed *original*. After considering briefly whether copying is possible in the oral mode, this paper focuses on the question of what constitutes a copy in written language and, more specifically, quoting in academic writing. It concludes with a discussion of the importance of copying for processes of language change.

Keywords Original language use, plagiarism, copying in academic quotations, copying in language processing, copying and language change

Introduction

Instances of copying occur in many contexts, as is shown throughout this volume. As a consequence, it is understandable that speakers of any language should also talk and write about the process of copying as such. In the *British National Corpus*, a digital collection of 100 million English words, the word *copy* (as a noun or verb) can be found in contexts such as the following:

- **Objects**: *People were always coming into the shop looking for things to* ***copy****.* [A6E 864]
- **Movements**: *Cherzeel scored with a close* ***copy*** *of England's goal.* [A9H 529][1]
- **Organisms**: *It replicates itself, and each new* ***copy****, which is independent of the original, goes on to carry out the task for which the virus was designed.* [A5R 720]
- **Painting**: *On my final day at Berkeley, my fourth-form pupils presented me with some flowers and a small mounted* ***copy*** *of Claude Lorrain's 'Hagar and the Angel'.* [A0F 632]
- **Film**: *IT IS amazing how few critics seem to have been able to recognise Dennis Potter's* Blackeyes *for what it is, just another Anglo-Australian* ***copy*** *of* Neighbours*.* [AA9 46]
- **Music**: *Consequently, a mechanical royalty arises each time a record company makes a* ***copy*** *of an album, cassette or compact disc.* [A6A 2422]

While this list is definitely not exhaustive, what can be noted is that, very frequently, the linguistic expression *copy* is employed to talk or write about copying in the domain of language, e.g. in the following examples from the *British National Corpus*:

- *Every document that goes through them* [photocopiers] *is recorded. Each* ***copy*** *is numbered and registered.* [A2X 426]
- *Her partner reads the magazine, too, and sometimes she lends a* ***copy*** *to a friend.* [A17 722]
- *Anne sent him a signed* ***copy*** *of her first book, Remembering Judi.* (sic) [ALJ 1967]

The corpus hits even suggest that the word *copy* most often refers to instances of language use. Note, however, that the individual, language-related uses of the word *copy* listed above still differ considerably. While the first is a synonym of *photocopy* (instances of which may also be effected of pictures), examples two and three do not refer to photocopies of books and magazines, respectively, but to individual examples from among a

1 For a discussion of copying with regard to a specific type of movement, namely dance, see Schwan, this volume.

series of printed items. This appears to be by far the most frequent use of the word form *copy* in the *British National Corpus*.

However, one must not forget that these are everyday uses of the word *copy* in language. In dictionaries of linguistic terminology, such as Brown's 2006 *Encyclopedia of Language and Linguistics*, the term *copy* is not listed as an entry. Instead, the word *copy* (both as a noun and as a verb) may occasionally appear in linguistic texts in its general usage. It is then used to refer to reproduced texts, and particularly manuscripts that were reproduced by handwriting, e.g. by monks during the Middle Ages.[2] In uses such as *copy-editing*, one finds the additional meaning of a manuscript being prepared for printing.[3] While the ideas represented by the term *copy* are usually expressed in linguistics by the terms *reproduction* or *imitation*, these are no central terms for the discipline, either. By contrast, actual copies of texts do play an important role in linguistics, insofar as they relate to the empirical study of language by means of corpora. Corpora are collections of texts that are usually deemed representative of a particular language or variety and can be searched for patterns. For instance, the *British National Corpus* was designed to represent the English language in general. Since many different aspects of language may be the focus of linguistic research (e.g. vocabulary, grammar, or spelling), corpus texts need to be faithful copies of the original texts. While they may be modified by adding information such as demographic details about authors, sentence numbers, part-of-speech tags etc., such coding is frequently restricted to a meta-level, so that the text as such remains unchanged.

Since *copy* is mainly used in its general meaning in linguistics, it is essential to define what is meant precisely by the term in the context of the present contribution. It makes sense to use the common meaning as the starting point for such a definition. The *Longman Dictionary of Contemporary English*, which uses a limited defining vocabulary, paraphrases the concept as "something that is made to be exactly like another thing," thereby capturing the essence of copying. According to this definition, a copy of an object is therefore expected to be an object,[4] a copy of a process is likewise a process,[5] and a copy of a linguistic entity is also an element of language.

Some concepts can be defined without having recourse to other, related, concepts. For instance, an ADULT can be defined as a person of a certain age without the need to explicitly refer to the concepts CHILD or TEENAGER. However, this is not possible for the idea of the copy, since it necessarily presupposes the idea of an original, upon which the copy is based. This aspect introduces a very strong chronological element. Consequently, it will be assumed in the following that it is only possible to produce a copy of an original with prior existence. This view differs from that maintained

2 Spencer and Howe 2001.
3 *Oxford English Dictionary*, s.v. "copy."
4 Cf. the examples in Stockhammer, this volume.
5 Cf. Schwan, this volume, on copying in dancing.

by Goodman and Latour, and by Lowe, who state that literary works of art do not go back to one original version, thereby implying that all versions of such a text have the same status.[6] In the same vein, the *Oxford English Dictionary* claims that individual examples (i.e. copies) of manuscripts or prints do not refer back to an original and uses the supporting argument that "the original draft is called the ***rough*** or ***foul*** copy."[7] According to Latour and Lowe, who discuss works of art, the relationship is even reversed, and the concept of originality presupposes the existence of copies: "No copies, no original."[8] Alternatively, one might wish to argue that this condition could be extended from the actual existence of copies to the quality of permitting to be copied—which would, in turn, raise the question of how similar something has to be in order to qualify as a copy rather than as a mere imitation or effigy, e.g. a small-scale model of a mountain which necessarily differs in size from the original. In a directly opposed view, however, one might want to argue that everything that is not a copy is an original. When the original undergoing copying in a secondary process is a copy itself, it would thus become a type of secondary original (e.g. when a student makes a copy of a worksheet he has received from his teacher, which is already a copy of an old original worksheet).

In linguistics, the term *original* is used in relation to texts which are translated into other languages.[9] It is also used to designate the steps that a word undergoes in the borrowing process: thus the English lexeme *chaos* was borrowed from Latin (as a so-called *proximate language*) but it is ultimately of Greek origin (its *original language*).[10]

Another aspect included in the simple definition above is the implication of an agent who deliberately attempts to achieve a likeness. This wording also leaves it open how similar the copy is in the end (e.g. due to the limitations discussed below). This is also true of the definition of *copy* found in the *Oxford English Dictionary*, one of whose meanings is "A transcript or reproduction of an original." To sum up what emerges as common ground in the various definitions of copying, a copy is an entity which is deliberately created with the aim of being exactly like an original.

The original and copy in language

Let us now focus on copying in language and begin by considering an example sentence:

(1) Linguistics is fun!

6 Goodman 1969, 114; Latour and Lowe 2009, 281. See also the discussion on "original" and "copy" with regard to texts in Graulund, this volume.
7 *Oxford English Dictionary*, s.v. "copy."
8 Latour and Lowe 2009, 278.
9 Lembersky 2012.
10 Cf. Hillebrand 1975, 224, and the *Oxford English Dictionary*, s.v. "chaos."

Compare this to the next example sentence:

(2) Linguistics is fun!

The two sentences are identical, with regard to both their form and their meaning. Since language consists of both form and meaning,[11] properties on both levels emerge as the prerequisite for copying in language. Sentences (1) and (2) would therefore seem to qualify as an instance of copying. If one of the two conditions does not apply, by contrast, we are definitely not dealing with a copy: for example, it is possible to reproduce the meaning of a linguistic entity relatively closely for a specific context by using various synonymous expressions. The utterance

(3) I find the study of language entertaining!

may thus fulfil the same pragmatic function as sentences (1) and (2) in certain contexts, but this is not an instance of copying, since it differs with regard to the formal side of the message. For the same reason, translations are not considered copies, since the target text differs in form and also, to a certain extent, in meaning from the original.[12]

Conversely, two sentences or utterances may be formally identical but have different meanings in the contexts in which they are used. Since the noun *plane* can refer either to an aircraft or to a surface, sentences (4) and (5) are different in spite of their formal identity, because the differing contexts (signaled by the sentences in parentheses) result in differing meanings.

(4) Look at that plane! (It is coming in our direction.)

(5) Look at that plane! (It intersects with line B.)

Sentence (5) can therefore not be considered a copy of sentence (4). However, this is a constructed example. In everyday language use, such instances that only superficially resemble copies are extremely unlikely to occur.

Another aspect that may be considered in discussions of copying related to language is how to treat co-occurring potential copies. This is the case in example sentences (1) and (2) above, but more commonly in cases such as

(6) Come in, come in!

11 Cf. e.g. de Saussure (1916/1959, 66–67) on the twofold nature of the linguistic sign.

12 E.g. because some linguistic associations or plays on words cannot be conserved due to the formal differences between languages.

and

(7) All right, all right!

which result in an intensification of the meaning.[13] When repetition occurs on the lexical level, that is inside a word, we speak of *reduplication*, e.g. in *ha ha* or *fifty-fifty*. Ghomeshi et al. present an interesting instance of this, namely contrastive focus reduplication, e.g. in

(8) That's not AUCKLAND–Auckland, is it?

This unusual juxtaposition "restricts the interpretation of the copied element to a 'real' or prototypical reading"[14]—in the example, to the famous city in New Zealand as against other places bearing the same name. Repetition and reduplication can thus be considered special instances of copying in language.

The first impulse in these cases is to claim that the original and the copy are used next to each other, just as one might want to argue that example sentence (2) is a copy of (1). However, the notion of copying in language is complicated by the twofold nature of language: one of the tenets of the school of structuralism is the necessary distinction between the levels of *langue* and *parole* in language. While *langue* represents the system of a language in the sense of an inventory of lexemes and a set of grammatical rules,[15] *parole* is defined as language usage in concrete utterances.[16] This is a very important distinction, since it affects whether linguistic entities should be evaluated as originals or copies. If we consider sentence (1) in this new light, we find that it consists of three words, namely *linguistics*, *is*, and *fun*. All of these words are established in the community of speakers of English and they are consequently part of the English *langue*. The question that now emerges is whether to consider each instance in which a language user uses a word like *linguistics* as a copy. After all, one might argue that each of these entities in the *parole* is the realisation of an element of the *langue* in actual usage. *Langue* and *parole* are related to each other via the minds of individual speakers. However, in view of the difference between the two systems, it is argued here that this is not an instance of copying but rather of some other transformative process, comparable to the way in which a spoken and a written sentence with the same content are not copies of each other. By contrast, all concrete realisations of a linguistic item in the same modality—e.g. *linguistics* in sentences (1) and

13 Ghomeshi et al. 2004, 318.
14 Ghomeshi et al. 2004, 307.
15 De Saussure (1916/1959, 9) defines *langue* as "both a social product of the faculty of speech and a collection of necessary conventions that have been adopted by a social body to permit individuals to exercise that faculty."
16 De Saussure (1916/1959, 13) characterizes *parole* as the "executive side" of language.

(2)—could be considered copies. Yet, strictly speaking, *linguistics* in sentence (2) is not a copy of *linguistics* in sentence (1). Rather, they share a common matrix in the *langue* that is in a certain way comparable to casting, in which a (quasi-)identical series of objects is produced by using the same mould.[17] While one might be tempted to argue that there is a difference between words and cast objects in that the production of a mould usually requires the prior existence of a positive model from which the negative mould is created, one should not forget that new words are used by an individual speaker on a singular occasion in that speaker's *parole*, and that it is through their use by other speakers that new words spread and finally enter the *langue* of a language.[18]

Copying as we shall consider it in the following sections thus primarily affects the *parole*. Note that some larger linguistic entities such as *Good morning* or *I am sorry* can be considered fixed expressions that are stored as single units in language users' minds, in spite of the fact that they can be analysed grammatically (e.g. *Good morning* as a noun phrase with a premodifying adjective). These complex units are therefore treated in a similar way to individual lexemes. Example sentence (1), by contrast, forms no common chunk but is the result of combining language-system-inherent words and rules (e.g. a singular verb form with the superficially plural noun *linguistics*). In this sense, sentence (1) is new and could be considered an original, and sentence (2) would be a copy of it. It is therefore arguable that linguistic copying, in the strictest sense of the word, only takes place if complex entities in the *parole* that are not single-unit entities in the *langue* (or at least not yet) are reproduced in parole again. It is in this sense that one may argue against Goodman,[19] and Latour and Lowe,[20] to claim that even prints of literary works go back to an original. That the copied entities usually follow the rules of the *langue* is a common correlation but not necessary, since it would also be possible to imitate nonsense words in the *parole*, or ungrammatical sentences contradicting the system of the *langue*.

However, the classification of a linguistic entity such as *Linguistics is fun!* as an original needs to be called into question if—as is the case here—it has been used before by the same author on another occasion, e.g. in a script prepared for teaching. Yet even if one were to retrieve the text in which a particular author wrote *Linguistics is fun!* for the very first time, one could only be certain to have come across the original with regard to production but not with regard to perception. After all, the author may have

17 Cf. also Goodman (1969, 112–113), who describes a similar relationship regarding music when stating that "all accurate copies [...] are equally genuine instances of the score" and calls music an *allographic* system contrasting with *autographic* painting, in which "even the most exact duplication" does not "count as genuine" (113).
18 Cf. de Saussure 1916/1959, 9.
19 Goodman 1969.
20 Latour and Lowe 2009.

read the sentence *Linguistics is fun!* before that occasion, without being aware of it now or then. This raises the question whether one can ever accept a linguistic utterance as an original, since any utterance may reproduce some other utterance that the language user or observer is currently unaware of. Therefore, in the context of this contribution, the independent production of the same novel utterance by different language users is defined as resulting in two originals, and one may even extend the notion of linguistic originality to the production of a linguistic utterance that is identical to another one produced previously, by the same language user, if that person is unaware of the earlier utterance.

The production of language will always require, at least to a certain extent, the reproduction of previously-used linguistic material (e.g. sounds, words, and grammatical rules) which, in their new combination, yield new instances of *parole*. Without potential reproducibility, a code such as language cannot function.[21] *Linguistics is fun!* is a relatively short utterance that observes the usual grammatical rules of English. As a consequence, the likelihood that one or more language users may have produced that sentence before is very high. In the past, it was assumed that language worked according to a slot-and-filler model, in which any grammatically suitable word may be inserted into the slots provided by syntax.[22] Modern linguistics, by contrast, recognizes the importance of chunks and prefabricated units,[23] and the open-choice principle has been largely replaced by an idiom principle.[24] As a consequence, one may expect that, in any given utterance, a certain proportion of chunks will be identical to chunks that have occurred in other utterances by the same speaker or by other speakers. While this would seem to suggest a very large degree of overlap between individual texts or utterances, there are also various factors that contribute to their potential uniqueness:

- increased length of the text or utterance
- infrequent vocabulary
- unusual collocations.

The longer a sentence (in terms of the number of words it comprises), the less likely it is to be produced as an original by different speakers on different occasions, e.g.

(9) Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather on which a mirror and a razor lay crossed.

21 Cf. Derrida 1999, 333.
22 Cf. the critical discussion in Sinclair 1991, 109.
23 Cf. Granger and Paquot 2008; Erman and Warren 2000.
24 Following Sinclair 1991, 110–115.

This sentence, the first from James Joyce's novel *Ulysses*, also fulfills the other two criteria: if a sentence contains infrequent words (e.g. *stairhead* with 14 hits in the 100-million-word *British National Corpus*), it also becomes more distinctive. The uniqueness of word combinations (so-called *collocations*) can be determined by using statistical measures such as *mutual information*, which compares the probability of encountering two lexemes in combination to the probability of encountering each individually in a corpus.[25] Thus *lie* is very frequent and *crossed* is relatively frequent, too, with 19,268 and 122 hits, respectively, in the *British National Corpus*, but their combination is quite unusual. A Google search for the sequence "lay crossed" on 7 November, 2016 yielded 44,500 hits, of which the majority of those viewed were quotations of Joyce's sentence. The more factors coincide, the more distinctive and original a text or utterance becomes and, as a consequence, the more likely its precise reproduction is to be a deliberate copy rather than mere coincidence. Where the boundaries should be drawn is, however, a matter of gradience. This is of particular importance in the detection of plagiarism. In their test of various types of plagiarism detection software, Weber-Wulff et al. find that some systems classify original texts as plagiarism "if the text uses many common phrases and the system reacts to four or five words in sequence as being plagiarism without examining a wider context."[26] Such false positives were returned, for example, for a text containing the sequences *Stieg Larsson was born in 1954* as well as *The rest of his childhood he lived* and *For the next birthday he got a*. Since most researchers would presumably agree that this is not original language use deserving to be protected by copyright, more refined algorithms are desirable to help decide what constitutes originals and copies in language, to support teachers in the detection of plagiarism.

Everything said about copying so far has been aimed at describing how it relates to language in general. In the next step, we will focus on modality-specific particularities. In principle, it is possible to copy orally, but this is less likely to occur than it is in written language, since long sequences in particular are harder to memorize and reproduce, due to the transitory nature of speech. According to Bakhtin, a large proportion of language users' communication consists of the re-telling of the text of others in their own words, but the formal modification combined with a certain semantic difference prevents such instances from being actual copies.[27] In the oral tradition, copying in the strictest sense mainly affects short utterances, such as slogans. The copying of whole texts in an oral tradition is most likely to occur with rhymed texts such as songs or poems, because these are easier to remember and reproduce verbatim.[28] However, even in these

25 Cf. Church and Hanks 1989.
26 Weber-Wulff et al. 2013.
27 Bakhtin 1981, 338–341.
28 The situation described in Ray Bradbury's novel *Fahrenheit 451*, in which individuals memorize whole books, is therefore highly uncommon. Another interesting situation is provided by acting: stage actors usually attempt to

cases, one may ask how similar two texts need to be in order to count as legitimate copies. In the oral reproduction of text, identity can hardly ever be achieved, since many features vary for the following reasons:

- Each speaker has their own voice. An identical spoken copy would therefore need to be produced by the same speaker. Furthermore, the speaker's voice quality may vary due to age, illness, etc.
- Every time a linguistic unit is reproduced, even by the same speaker, a difference in loudness, speed, intonation, accentuation, or structuring by means of pauses may occur.
- The perceived vocal quality also depends on the situation where the speaking takes place, e.g. in a small room vs. a staircase.

We may therefore conclude that a copy may be quasi-identical on the level of the linguistic system, but that this can hardly ever be achieved on the level of concrete realization in oral speech. Every time an artist recites a poem, the result will slightly differ. The best auditory copies in this sense are produced by means of sound recording and subsequent copying to other data carriers. Note, however, that not even digital copies are one hundred percent identical: while the acoustic realisations of the same recording through the same loudspeaker in the same context will presumably be indistinguishable from each other (in contrast to gramophones, which produce a variety of accompanying noises), digital copies are at least distinct on the meta-level. Every time a file is copied, information on the copying process is encoded in the file's details, because the time at which a new file is created in the copying process is part of the dataset. Each copy is thus unique in a certain aspect, like a banknote with its serial number.[29] One may, however, argue that this aspect of the copy is irrelevant to its functionality and only applies on the meta-level.

While it would seem that it is almost impossible to copy longer stretches of spoken language with exactitude, written texts seem to lend themselves far more readily to copying. Indeed, two exemplars of an article in two hard copies of the same newspaper are virtually indistinguishable from each other. However, even written texts may differ with regard to a number of fine nuances:

- Handwritten texts pose the same problems as noted above: two identical sentences written by the same person are hardly ever one hundred percent identical formally. They will differ with regard to the material (the size of the paper, the writing instrument, the ink color, etc.), the

reproduce the scripted version of a play word for word—thereby copying between modalities if one were to recognize this as copying—but note that their interpretation is considered an important aspect of the acting process.

29 Cf. Schröter, this volume.

amount of pressure used in writing, or the size of the letters. No two signatures are identical, either.
- The medium will even change the visual impression of a typed text. There is thus a difference between a digital text on the author's computer (even as a PDF) and the same text as a printed version using paper, toner, and printer ink.

Most of the time, however, most language users will happily perceive printed texts as identical copies of each other. On a general level, one may therefore distinguish between an absolutely faithful copy and a functionally faithful copy. While absolute identity between various copies of a text can arguably never be achieved in language because there will always be some differences (even on the atomic level of the paper used for printing), functionally faithful copies abound, particularly in writing. These are the copies Goodman has in mind when he states that "correct copies" of literary works are based on "sameness of spelling" and consist of "exact correspondence as sequences of letters, spaces, and punctuation marks," further asserting that "any sequence [...] that so corresponds to a correct copy is itself correct, and nothing is more the original work than is such a correct copy."[30]

Copying in academic quotations

Printed copies of a text are made in order to expand its potential readership. If a whole text is copied very frequently, this can be regarded as an indication of its popularity and/or importance. In academia, a subtype of copying is frequently observable within texts: *quoting* does not involve the copying of complete texts but merely of passages from texts by other researchers. The beginning and ending of a quoted passage are usually indicated by the use of quotation marks.[31] Quoting, which is usually defined as a verbatim reproduction of an original text, contrasts with paraphrasing, which implies the use of the paraphrasing author's own words to convey the paraphrased author's idea(s). Both in quoting and in paraphrasing, the source is indicated in order to make it clear that a specific idea is attributable to another author. In quoting, this extends to the wording as well. The reason for quoting rather than paraphrasing another author may be respect for that particular author, as well as the feeling that a specific wording is the best possible way to express an idea. Other reasons may

30 In this sense, copies of literary texts differ from copies of paintings, whose correctness in copying is more difficult to determine due to the fact that the properties of pictures cannot be broken down into discrete features so easily (see Goodman 1969, 115–116).

31 This use of inverted commas in order to mark stretches of text as quotations is a relatively recent convention (Moore 2011, 1).

be the wish to state precisely a view that contradicts the author's,[32] or the necessity to provide the reader with a statement that will be analyzed in more detail (as when various definitions of a concept are contrasted with each other).[33] Since quotations are directly attributed to other authors, it is the quoter's responsibility not to modify the original text in any way.[34] Otherwise, the authors of the original text might be presumed responsible for the distorted ideas in supposed quotations of their text. As a consequence, it would seem that there is an opposition between identical reproductions—i.e. quotations—at the very extreme end of the scale of copying and a large degree of variation in the extent to which the original wording and order of ideas are modified in paraphrases, e.g. through replacement with synonyms, rearrangement of the clause elements, or the passivization of active sentences.[35]

Yet contrary to expectations, quotations in academic texts are not necessarily one hundred percent identical to the original passage. Not all features of the original are conserved, and in some cases, modifications are even required. The following collection of noteworthy aspects of quotes permits determining what is necessary in order to accept a text as a quasi-identical copy of another text in the context of academic quotations. It thus provides information on what constitutes the essence of language in the copying process:

- Any quotation necessarily conserves the meaning of a quoted passage. This is usually reflected in the reproduction of its letters, numbers, symbols, punctuation marks, spacing, capitalisation, small capitals, bold print, italics, and line breaks.
- By contrast, a quotation does not usually conserve the font type. A text in Arial may be quoted in Times New Roman without the need to mention this change. Font type is thus regarded as part of the standard formal background of a text. However, if an original text were to use various fonts in order to transmit a message, e.g. in a (constructed) original sentence such as

(10) Short passages are easier to read if typefaces without serifs are used (e.g. in *This is a serif typeface* as opposed to *This is a sans serif typeface*).

32 Booth, Colomb, and Williams 2008, 97.
33 Cf. Schneider 2015 for a discussion of quoting as a positioning strategy, e.g. in order to express criticism.
34 Achtert and Gibaldi (1985, 71) demand that "in general, a quotation—whether a word, phrase, sentence, or more—should correspond exactly to its source in spelling, capitalization, and interior punctuation."
35 See also Booth, Colomb, and Williams 2008, 192–195.

one would expect a stricter observance of formal aspects in quoting than usual—at least for the self-referential text passage printed in the font type under consideration.

- Similarly, font size is disregarded in quoting. Once again, the standard of the original is legitimately captured by any standard of the copy. However, in word-processed texts, longer quotations are often signalled by indentation and sometimes also by a smaller font size and spacing between the lines than is present in the remainder of the text,[36] e.g. in the following passage:

(11) The unity of the chain of letters between two spaces in solid compounds can be regarded as an indication of how strongly they belong together.[37] Thus Haiman writes that

> The distance between linguistic expressions may be an iconically motivated index of the conceptual distance between the terms or events which they denote. But the length of an utterance may also correspond to the extent to which it conveys new or unfamiliar information. Reduced form may thus be an economically motivated index of familiarity.[38]

To sum up, copied passages which are not framed by quotation marks are obligatorily modified in their formatting. If there is a marked difference in font size between parts of the quoted text, this is presumably conserved if it matters, e.g. in the case of self-referentiality. However, one aspect that is not usually conserved in quotations is the use of special initials. Thus, the edition of *Ulysses* used for the example quoted above actually extends the first letter of the first word <S> over two lines and spells the remaining sequence <TATELY> (in capital letters) as a continuation of that special style. Quoting this precise formatting would be very difficult. Since this formatting convention applies to all beginnings of all the parts into which the book is divided, one may also interpret this as a purely editorial decision, and one therefore lying outside the "original" text by Joyce. This supposition is supported by the numerous quotations of the sentence online, of which only a minute proportion uses capitalisation (and none the extra-large initial letter), possibly because they refer to some other version, but more probably in order to arrive at format that is easier to produce technically.

36 These are defined in the 2009 APA style guide as quotations consisting of at least 40 words.
37 Cf. Erben 2007, 112.
38 Haiman 1983, 781.

- Line length plays no role in the production of a faithful quotation, either. Usually,[39] copied passages are quoted as flow text.[40] This raises the question of how to proceed with the end-of-line hyphenation present in the original text. Interestingly, the *MLA Style Manual*, which demands that "the internal punctuation of a quotation must remain intact,"[41] does not mention this aspect at all. In the majority of cases, the end-of-line hyphen is simply deleted, and a hyphenated original like *wa-ter* becomes *water* in the copy. Keeping the hyphen in a place other than at the end of the line in the target text would result in an unusual sequence and thus an unintentional variation from the original, e.g. in the unconventionally spelled

(12) The isle was surrounded by wa-ter.

The opposite strategy of conserving the hyphen's place at the end of a line would require the insertion of an additional paragraph break in most text layouts. However, this would disrupt the usual formatting of the quoting text, and it is possibly for that reason that this strategy seems to be highly unusual. Copying may therefore require the deletion of a sign (i.e. the hyphen) in order to conserve the intended form of the original text and thus to produce a functionally faithful copy. This task is made more difficult by the fact that some texts contain prefixations such as *co-operate* or compounds such as *bitter-sweet*, which are frequently but not always hyphenated.[42] In these cases, the copyist has to guess the original spelling preferred by the author in order to observe the principle of formal reproduction. In linguistics, this aspect is of particular importance to compilers of corpora, who must copy long passages from texts. In a forthcoming publication, I have found references to copying strategies in various manuals from the Brown corpus family: one strategy to resolve such ambiguous cases is to use the spelling of other instances of the same compound in the same text; another is to use an authoritative reference work.[43] Note that the copy-pasting of texts from PDF files into MS Word files deletes end-of-line hyphens by default. This is problematic if the end-of-line hyphen in the original coincides with the hyphen of a compound that is almost exclusively hyphenated in linguistic usage (e.g. *hard-working*).[44] If the hyphen

39 Note, however, that up to three lines from poems (a genre in which the ends of lines play an important role) may be quoted by representing the line breaks by means of slashes with spaces on each side (Achtert and Gibaldi 1985, 73) and that turnover lines in quoted poems (which are too long for the format of the quoting text's lines) are indicated by using the code [t/o] (Achtert and Gibaldi 1985, 75).
40 The idea that this is no new phenomenon is supported by Wetzel (1981, 28–29), according to whom the scribes of earlier times did not reproduce original line length, either.
41 Achtert and Gibaldi 1985, 80.
42 Cf. Sanchez-Stockhammer, forthcoming.
43 Sanchez-Stockhammer, forthcoming.
44 Cf. Sanchez-Stockhammer, forthcoming.

is deleted by default, the result (in the example, *hardworking*) will most certainly contradict the spelling intended by the author of the original text. As a consequence, end-of-line hyphens should only be deleted after individual consideration, and one cannot always be certain of having captured the intention of the original author, because some compounds or prefixations permit variation.

- If a feature of an original text cannot be preserved in a quotation due to some restriction of the medium (e.g. in typewriting or handwriting), there are usually conventions which permit the creation of an alternative that counts as a legitimate copy:

    — Italicisation may be represented by underlining,[45] bold print by framing a word or expression with asterisks (e.g. in *This is *great*!*), and small caps by using normal capitals.
    — Herbst and Klotz, for example (whose work is printed in black-and-white), indicate the use of color in quoted dictionary headwords by employing underlining accompanied by an explanation in parentheses.[46]
    — If an en dash <–> or em dash <—> is unavailable, either can be represented with a hyphen.[47] Note, however, that the replacement of an em dash (which is surrounded by letters) by an en dash or a simple hyphen may result in the insertion of spaces so as to distinguish it formally from a word-internal hyphen (which is surrounded by letters).[48] This needs to be considered in the analysis of text-only corpora, since attaching a hyphen to the end of a word may prevent it from being found by a search pattern that utilizes standard spacing.
    — Yet another issue to consider is the use of diacritics such as the cedilla <ç> or the hacek <ě>. Since these are distinctive and potentially lead to differences in meaning, their omission in the target text would reduce the accuracy of a copy. As a consequence, diacritics in typed texts have often been manually added to the paper copy of a quotation. Corpora frequently use special codes to represent diacritics.

- Since quotations need to be verbatim copies of the original text, this means that mistakes also have to be copied. However, it is possible to add the commentary [sic]—Latin for "thus"—in square brackets,[49] in order to show that a spelling mistake was not inserted by the copyist (which would be the reader's usual assumption). Such brackets

45 Cf. Achtert and Gibaldi 1985, 78.
46 Herbst and Klotz 2003.
47 Cf. Huddleston and Pullum 2002, 1725–1726.
48 Cf. Sanchez-Stockhammer, forthcoming.
49 Achtert and Gibaldi 1985, 78.

can also be used to mark explanatory additions (e.g. who a personal pronoun in a passage refers to), omissions (signalled by [...][50]), or modifications (e.g. if capitals are changed to lower case or inflections are adapted in number or tense). Consequently, a legitimate, functionally faithful copy may deviate from the original as long as all changes are indicated in square brackets. These modifications on the meta-level do not make the quotation a paraphrase but simply a modified quotation. However, style guides differ in the extent to which they accept unmarked changes in a quotation: thus the American Psychological Association's style guide permits the conversion of the first letter of the first word in a quotation to upper or lower case and even states that "[t]he punctuation mark at the end of a sentence may be changed to fit the syntax" and that "[s]ingle quotation marks may be changed to double quotation marks and vice versa."[51] The *MLA Style Manual*, by contrast, demands the modification of capitalisation in square brackets, as exemplified in the preceding two quotations.[52] A problematic situation occurs if an original passage already contains square brackets. In such cases, a copyist may feel the urge to add a comment in another pair of square brackets in order to state that the first pair of brackets was already present in the original.

- Quotation marks also provide an interesting case. If a passage containing quotation marks is quoted, the usual convention is that the type of quotation mark inside the copied passage contrasts with the type of quotation mark used in the quoting text: if the meta-text uses double quotation marks, all quotation marks inside the quoted passage are rendered with single quotation marks, and if the meta-text uses single quotation marks, all quotation marks inside the quoted passage are rendered with double quotation marks.[53] This prevents confusion as to where the quoted passage begins and ends. When the quoting and the quoted text use the same type of quotation mark, this requires an adaptation of the quotation marks; if they use different types anyway, there is no conflict. An interesting question in this context is how to treat foreign quotation marks (e.g. from German or French original texts) in English texts. In contrast to English quotation marks < “ > < ” >, the opening German quotation mark < „ > is at the bottom, while the closing mark < “ > corresponds to the English opening mark, and French quotation marks look completely different < « > < » >. As a consequence,

50 Note, however, that neither the APA style guide (2009, 176) nor Achtert and Gibaldi (1985, 76) demand the use of brackets around omission points, which means that the readers will have to infer that these represent an alteration by the author of the quoting text and are not part of the quote itself.
51 American Psychological Association 2009, 176.
52 Achtert and Gibaldi 1985, 79.
53 Cf. Achtert and Gibaldi (1985, 80), who only consider their own standard case of double quotation marks for the main quotation and single quotation marks for the embedded quotation.

these foreign quotation marks could remain unchanged—but most of the time, they will presumably be replaced with the standard quotation marks of the quoting text.

All of the above examples suggest that a copy of a linguistic utterance in the highly formal context of academic quoting may differ in certain ways from the original. Paradoxically, in some cases, the conservation of all original features—i.e. an absolutely faithful copy—would even result in a deterioration of the copy's functionality (e.g. in the case of end-of-line hyphenation).

The role of copying in language change

We have seen above that copying in language rarely leads to identical copies. At the same time, it is commonly recognized that variation in language represents the basis for linguistic change.[54] This raises the question of to what extent copying—and particularly imperfect copying—contributes to linguistic change.

A certain proportion of the variation observed in copies is the result of reduced cognitive ability (e.g. due to tiredness or inattentiveness). In children and language learners, failing memory regarding the precise form of an expression that has not yet been mastered completely may also play a role.[55] The proportion of language use affected by these obstacles must, however, not be overestimated.

According to Lass, "language transmission is replication" and language as such is "a replicating information system," in which "variance is 'copying error.'"[56] He states that some of the errors that "creep into the replication process" may be "stabilized by selection," while others are not.[57] However, the concept of error presupposes the existence of an alternative which is evaluated as being more correct (or at least as more appropriate) in a particular context than the variant used. Since the question of who can legitimately make such evaluative judgments is highly controversial, linguistic variation is frequently discussed on more neutral grounds in linguistics. In many cases, variation in the copy is deliberate, e.g. because the copyist wishes to make a particular point or to introduce his or her own view. This is for example the case when the title of a famous book, song, film, etc. is modified to make a pun.[58] As we have seen above, most instances in

54 Cf. e.g. Holmes 2008, 205–206.
55 This finding was supported by an experiment by Kirby, Cornish, and Smith (2008), in which the imperfect reproduction of artificial linguistic stimuli was found to lead to systematic changes.
56 Lass 1997, 112–113.
57 Lass 1997, 112.
58 See Bolte's 2005 volume *Paradigms Lost*, whose title alludes to John Milton's classic *Paradise Lost.*

which language users refer to other language users' texts or utterances are therefore not copies in the strictest sense, but comprise form-meaning modifications of some kind and would therefore rather be classified as paraphrases, according to the definition used here.

Since language change on an abstract level presupposes change in the language of individual speakers, it makes sense to consider the role of copying in language change by adopting a cognitive linguistic perspective. One might be tempted to assume that the storage of identical copies of words (e.g. regarding denotative meaning, spelling, and pronunciation) in all speakers' mental lexicons is a prerequisite for successful communication. However, this is not the case: not only will the words in individual language users' minds differ with regard to the unique biographical experiences with which they are associated, but in a commonly-used analogy, the mind is compared to a corpus in which all previously encountered language is stored in some way or another.[59] Since every language user has experienced a unique combination of linguistic input, individual mental corpora must therefore necessarily differ from each other as well. It is only in the sense of a shared common ground for each lexeme that we can speak of some kind of copy (in the widest sense) existing in the minds of different speakers at all.

By contrast, the concept of copying also plays another role: according to Bybee's exemplar- and usage-based linguistic model of *emergence*, "certain simple properties of a substantive nature, when applied repeatedly, create structure."[60] The frequency of usage of linguistic expressions encountered in linguistic input—i.e. the occurrence of linguistic copies—thus shapes the mental corpora of the language users. Since individual language users base their own linguistic decisions (e.g. whether to use *whom* or *who*) on the frequency of linguistic phenomena in their mental corpus, the frequently repeated perception of copies in the past will consequently influence the future linguistic behaviour of individual speakers.

The frequency-dependence of change is also true for language in general. This is comparable (but not identical) to repeated photocopying: if an original text is inserted into a photocopier over and over again, the material on which the text is printed will deteriorate in the course of time, and this will influence future photocopies of the text, which will then differ from earlier photocopies in that they will also reproduce an increasing number of smudges, creases, etc. When transferring this principle to language, we can observe that if a word is frequently repeated

59 Cf. Taylor (2005, 3), who also notes (13) that there are differences regarding the amount of detail presumably stored in memory compared to a linguistic corpus (e.g. regarding the context in which linguistic forms were encountered), the possibly different format (linear text vs. an assumed hypertext-like format in memory), and the differing temporal dynamics (involving the inclusion of new linguistic forms and the potential decay of memory traces compared to a stable conventional corpus).

60 Bybee 2003, 3.

in discourse, this leads to a reduction of its form in the pronunciation.[61] In some cases, such as *evening* (/ˈiːvnɪŋ/), this has already resulted in a standard pronunciation that drops the second syllable.[62] Similarly, English uses so-called *weak forms* for the majority of frequent grammatical words (i.e. pronouns, prepositions etc.):[63] thus the third person singular verb form *has* is hardly ever pronounced /hæz/ with a full vowel (with the exception of when it is used as a full verb, e.g. in *He has a car*). Usually, it is pronounced as /həz/, /əz/, or even /z/ or /s/, e.g. in *He has been here*.[64] This change is even reflected in the use of the contracted spelling *He's been here*.[65]

While it is possible to observe instances of imperfect copying in language use, the influence of copying errors in shaping language should not be overestimated. Instead, it is copying in the sense of repetition (which results in the increased frequency of use of linguistic expressions) that plays the most important role as a transformative power. This is also true of new language uses, i.e. the basis for linguistic change, which need not be imperfect copying but may rather represent alternative creations by different language users.

Conclusion

While this contribution is limited to observations concerning English, one may assume that many of the aspects touched upon are not language-specific and can be transferred to other languages without claiming universal status *a priori*.

To sum up, copying in language needs to consider two levels: that of the linguistic system and that of the concrete realization of language in utterances. This might distinguish copying in language from copying in at least some other systems. Depending on the size of the entities under consideration and the desired level of similarity, we find that, while functionally faithful copying in language is extremely frequent (e.g. if we consider that almost all words in a text have been used before in a language), it is practically impossible to create an identical copy of language use, due to situation-dependent variation, particularly in the spoken reproduction of longer passages.

Academic quoting represents a special case of copying in language. While absolute identity between the original passage and its reproduction

61 Bybee 2003, 8–9; 58.
62 Cf. Wells 2008 at *evening*.
63 Cf. Eckert and Barry 2005, 215–216.
64 Wells 2008 at *has*.
65 Interestingly, for levels of language other than pronunciation, the failure to copy a particular expression frequently enough may also lead to its changing: thus infrequent irregular verbs, such as *weep/wept*, are more likely to be regularised (to *weeped*) than frequent irregular forms (e.g. *keep/kept*), which are more stable due to their stronger representation in memory (Bybee 2006, 715).

is explicitly demanded, the presence of certain features in a text—particularly punctuation marks—actually requires deviations from the original, in order to produce a functionally faithful copy.

Even if the influence of copying errors on language change should not be overestimated, we can still observe that copying as such plays an important role in language change: since repeated exposure to a linguistic pattern in communication incites language users to modify their own linguistic production, we can conclude that copying in language has strong transformative potential indeed.

References

Achtert, Walter S., and Joseph Gibaldi. 1985. *The MLA Style Manual*. New York: The Modern Language Association of America.

American Psychological Association. 2009. *Concise Rules of APA Style*. 6th ed. Washington: American Psychological Association.

Bakhtin, Mikhail. 1981. *The Dialogic Imagination: Four Essays*. Austin: University of Texas Press.

Bolte, Henning. 2005. *Paradigms Lost*. Duisburg: OBST.

Booth, Wayne C., Gregory G. Colomb, and Joseph M. Williams. 2008. *The Craft of Research*. 3rd ed. Chicago: The University of Chicago Press.

Bradbury, Ray. 1984. *Fahrenheit Four Hundred and Fifty-One*. New York: Ballentine Books.

British National Corpus. Accessed August 4, 2014. http://bncweb.lancs.ac.uk.

Brown, Edward K. 2006. *Encyclopedia of Language and Linguistics*. Amsterdam: Elsevier.

Bybee, Joan. 2003. *Phonology and Language Use*. 2nd ed. Cambridge: Cambridge University Press.

———. 2006. "From Usage to Grammar: The Mind's Response to Repetition." *Language* 82 (4): 711–732.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.

Derrida, Jacques. 1999. *Randgänge der Philosophie*. Translated by Gerhard Ahrens. 2nd ed. Vienna: Passagen-Verlag.

Eckert, Hartwig, and William Barry. 2005. *The Phonetics and Phonology of English Pronunciation*. Trier: wvt.

Erben, Johannes. 2007. "Die Tendenz zum Aufbau mehrgliedriger Wörter im Deutschen und Versuche, die wortinterne Gliederung lautlich oder graphisch zu verdeutlichen." *Zeitschrift für Deutsche Philologie* 126 (1): 111–118.

Erman, Britt, and Beatrice Warren. 2000. "The Idiom Principle and the Open-Choice Principle." *Text* 20 (1): 29–62.

Ghomeshi, Jila, Ray Jackendoff, Nicole Rosen, and Kevin Russell. 2004. "Contrastive Focus Reduplication in English; The Salad-Salad Paper." *Natural Language & Linguistic Theory* 22 (2): 307–357.

Goodman, Nelson. 1969. *Languages of Art: An Approach to a Theory of Symbols*. London: Oxford University Press.

Granger, Sylviane, and Magali Paquot. 2008. "Disentangling the Phraseological Web." In *Phraseology: An Interdisciplinary Perspective*, edited by Sylviane Granger and Fanny Meunier, 27–49. Amsterdam: Benjamins.

Haiman, John. 1983. "Iconic and Economic Motivation." *Language* 59 (4): 781–819.

Herbst, Thomas, and Michael Klotz. 2003. *Lexikografie*. Paderborn: Schöningh.

Hillebrand, Ulrich. 1975. "Chronologische und etymologische Untersuchungen zum französischen Wortbestand innerhalb der englischen Sprache." PhD diss., University of Münster.

Holmes, Janet. 2008. *An Introduction to Sociolinguistics*. Harlow: Pearson.

Huddleston, Rodney, and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Joyce, James. (1922) 1997. *Ulysses*. London: Picador.

Kirby, Simon, Hannah Cornish, and Kenny Smith. 2008. "Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language." *National Academy of Sciences of the United States of America (PNAS)* 105 (31): 10681–10686.

Lass, Roger. 1997. *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

Latour, Bruno, and Adam Lowe. 2011. "The Migration of the Aura, or How to Explore the Original through Its Facsimiles." In *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*, edited by Thomas Bartscherer, 275–297. Chicago: The University of Chicago Press.

Lembersky, Gennadi. 2012. "Language Models for Machine Translation: Original vs. Translated Texts." *Computational Linguistics* 38 (4): 799–825.

Longman Dictionary of Contemporary English. 2009. 5th ed. Harlow: Pearson.

Milton, John. 1667. *Paradise Lost: A Poem Written in Ten Books*. London: Peter Parker, Robert Boulter & Matthias Walker.

Moore, Colette. 2011. *Quoting Speech in Early English*. Cambridge: Cambridge University Press.

Oxford English Dictionary. 2009. 2nd ed. Oxford: Oxford University Press. CD-ROM, 4.0.0.3.

Sanchez-Stockhammer, Christina. Forthcoming. *English Compounds and Their Spelling.* Studies in English Language. Cambridge: Cambridge University Press.

Saussure, Ferdinand de. (1916) 1959. *Course in General Linguistics.* Translated by Wade Baskin. New York: Philosophical Library.

Schneider, Klaus P. 2015. "Manufacturing Credibility: Academic Quoting Across Cultures." In *The Pragmatics of Quoting Now and Then*, edited by Jenny Arendholz, Wolfram Bublitz, and Monika Kirner-Ludwig, 209–227. Berlin: De Gruyter.

Sinclair, John. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Spencer, Matthew, and Christopher J. Howe. 2001. "Estimating Distances Between Manuscripts Based on Copying Errors." *Literary & Linguistic Computing* 16 (4): 467–484.

Taylor, John R. 2005. *The Mental Corpus: How Language is Represented in the Mind*. Oxford: Oxford University Press.

Weber-Wulff, Debora, Christopher Möller, Jannis Touras, and Elin Zincke. 2013. *Plagiarism Detection Software Test 2013*. Accessed January 25, 2014. http://plagiat.htw-berlin.de/software-en/test2013/report-2013/.

Wells, John C. 2008. *Longman Pronunciation Dictionary*. Harlow: Pearson.

Wetzel, Claus-Dieter. 1981. *Die Worttrennung am Zeilenende in altenglischen Handschriften.* Frankfurt am Main: Peter Lang.