CERLIS Series
Volume 4

Cécile Desoutter, Dorothee Heller & Michele Sala (eds)

Corpora in specialized communication
Korpora in der Fachkommunikation
Les corpus dans la communication spécialisée

CELSB
Bergamo

CERLIS SERIES Vol. 4

CERLIS
Centro di Ricerca sui Linguaggi Specialistici
Research Centre on Languages for Specific Purposes
University of Bergamo
www.unibg.it/cerlis

CORPORA IN SPECIALIZED COMMUNICATION
KORPORA IN DER FACHKOMMUNIKATION
LES CORPUS DANS LA COMMUNICATION SPÉCIALISÉE
Cécile Desoutter, Dorothee Heller & Michele Sala (eds)

ISBN 978-88-89804-25-4

Published in Italy by CELSB Libreria Universitaria
Via Pignolo, 113 - 24121, Bergamo, Italy

Indice

I corpora in contesti accademici

I corpora in contesti pedagogici

I corpora in contesti legali

I corpora in contesti professionali

PATRIZIA ANESA

2. Avoiding Plagiarism and Self-plagiarism through the Use of Corpora[1]

Introduction

Plagiarism is an ancient phenomenon that seems to be present in a vast array of eras, regions, disciplines, and fields, and a plethora of tools and procedures have been developed in order to try to contrast it. Indeed, plagiarism assumes different forms and, subsequently, plagiarism detection encompasses a wide range of methodologies that may be applied to different fields in the professional, academic and educational world. In particular, two of the main research trends focus on free text plagiarism (for a survey see Maurer *et al.* 2006) and on source-code plagiarism detection (e.g. Arwin/Tahaghoghi 2006).

This chapter focuses on text plagiarism and in particular on written academic texts. As regards academic writing, the battle against plagiarism seems to become more and more complex as it is often stated that the advent of the Internet has multiplied the opportunities for plagiarism (Howard 2007), making it easier and faster. For instance, if we look at the simplest and most obvious form of plagiarism, i.e. copy-and-paste procedures, we know that they are adopted worldwide by myriads of students. Although this battle seems, to some extent, bound to fail, different tools and strategies are being implemented in order to detect plagiarism efficiently and effectively. Given the nature of the principal contemporary forms of

[1] I am deeply indebted to Giuseppe Parrinello for the precious help received in this study, especially as regards Section 3.

plagiarism, it is generally argued that these tools are aimed in particular at the detection of digital plagiarism. Traditional plagiarism and ‘digital’ plagiarism are no longer clearly distinguishable given the number of publications made available in digital format and of publications which are only available digitally. Consequently, digital forms of plagiarism will most likely continue to be the most widespread.

The first section of the paper offers a taxonomy of plagiarism, showing the different forms assumed by the phenomenon and its main characteristics. The second section is devoted to a discussion of the main approaches to plagiarism detection. Finally, the third section focuses on how corpora or other types of text collections may be used for plagiarism detection and research, and it also presents the potential uses of *PlagiarismFinder* (*PF*), a programme specifically developed for preliminary external detection.

1. Defining plagiarism

Plagiarism may be broadly defined as “the use of another author's words, research, or ideas without proper attribution and citation” (Stepchyshyn/Nelson 2007: 63). Therefore, it involves the inappropriate use of intellectual property (Bloch 2012: 1) and, from a legal perspective, it does not necessarily equal copyright infringement in that it may be based on the duplication of material that is not copyrighted. More precisely, Murray highlights that plagiarism refers to the “use or reuse of words or ideas without *acknowledgment*”, whereas copyright infringement is the “use or reuse of words or ideas without *permission*” (Murray 2008: 174).

Plagiarism depends significantly on cultural factors and a monolithic interpretation of plagiarism is not possible. Buranen points out that cultural differences are not the only factor to be used in defining plagiarism, but such differences inevitably contribute to the complexification of the problem (Buranen 1999: 65). Indeed, the boundaries between rephrasing, paraphrasing, citing, and plagiarizing are often unclear and different cultures may assign different levels of importance to originality and creativity (for a discussion of plagiarism in different eras and cultures see Anderson 1998).

As has been mentioned, plagiarism may concern different fields, countries and languages, and the phenomenon is so widespread that it is impossible to clearly identify which categories of people are more inclined to commit plagiarism. For instance, as regards academic writing, some studies have focused on the incidence of plagiarism in works written by native and non-native speakers (Campbell 1990, Pecorari 2003), but no incontrovertible differences seem to emerge.

Plagiarism is often investigated in connection with other phenomena such as the practice of ‘ghost writing’ (writing texts that are officially credited to another person) and the use of essay mills. They represent forms of intellectual cheating, but not necessarily of plagiarism, however condemnable these practices may be. For instance, a ghost writer does not generally make use of plagiarism in its original sense. Similarly, essay mills are often sources of plagiarism in that the same essay may be sold to several students, and, therefore, the same text is attributed to different people who present it as their own. However, the original text may be exempt from instances of plagiarism.

The concept of plagiarism is extremely protean and its realization may assume a vast variety of forms. Different parameters could be used in order to identify different types of plagiarism, among which I would suggest the following:

- intentional and unintentional (or accidental) plagiarism;
- other-plagiarism and self-plagiarism;
- exact, near-exact and concealed plagiarism.

The next section will discuss these phenomena in detail.

1.1. Intentionality

Aspiring to adopt criteria that allow us to discern objectively and incontrovertibly between intentional and unintentional plagiarism is a process that is somehow doomed to failure. However, situations in which authors do not clearly acknowledge other people's words or ideas with the obvious aim of presenting them as their own is to be seen as a deliberate act of plagiarism. In contrast, occasional mistakes and inaccuracies should be framed within a different category; not using quotations marks when the source is cited, or, conversely, not citing a source but presenting a quotation correctly may be examples of unintentional plagiarism. This category is often associated with less experienced writers. Borderline cases are of course numerous: for example, a free and inaccurate use of class notes may be seen as a form of unintentional plagiarism linked to lack of experience on the part of a student in assigning citations and indicating sources. Moreover, a student may have genuinely misunderstood that a certain statement pronounced in class was in fact a quotation.

1.2. Other- and self-plagiarism

Text plagiarism in its simplest terms refers to the practice of using somebody else's words as if they were our own. Self-plagiarism, however, may also occur. The American Psychological Association defines self-plagiarism as follows: "Whereas plagiarism refers to the practice of claiming credit for the words, ideas, and concepts of others, self-plagiarism refers to the practice of presenting one's own previously published work as though it were new" (APA 2010: 170).

It is clear that in academic writing, especially at a scholarly level, a specific topic is often dealt with by the same author on numerous occasions, given the high level of specialization scholars tend to reach. Research is intrinsically based on the advancement of knowledge and publications are often conceived as follow-ups to previous works. When reading different publications by the same author, it is, therefore, understandable that the same topic may be introduced in similar ways or that the literature overview may lack a high level of originality. As the APA underlines, however, the key distinction lies in the lack of acknowledgement regarding previously published works. Not citing the (more or less explicit) contribution that such studies make to the current publication is to be seen as plagiaristic practice. It is neither erroneous nor immoral to start from the same theoretical background or to use the same methodological approach to analyze different sets of data in order to answer a different research question. However, blatantly re-using the same sources, the same data, or the same wording without a clear reference to previously published work is a debatable praxis.

The expression ‘self-plagiarism’ itself may appear oxymoronic. The debate has often focused on its very nature and on whether or not it should be considered a crime. If we assume that the notion of plagiarism can be related to that of theft, is it plausible to assume the existence of a crime such as self-theft? Moreover, if we observe plagiarism from a philosophical perspective, we seem to be dealing with a kind of Sorites paradox[2]: what is the line that discerns exactly what plagiarism is and what isn’t? How many words, ideas, concepts or methods can one ‘steal’ from himself before it constitutes a crime?

Practices related to self-plagiarism may be labeled in different ways, from more condemning expressions such as ‘text recycling’ to more laxist and apologetic ones such as ‘text re-use’ (Clough/ Gaizauskas 2009) or ‘text self-borrowing’. Different phenomena, such as duplication, recycling or cryptomnesia, are included in or related to the idea of self-plagiarism and at times overlap. Duplicating or recycling part of one’s research may be considered as a form of self-plagiarism, especially when dealing with an exact duplication of theories, materials, data, findings, etc., or when they are recycled without acknowledging their previous use. However, practices based on considerable modification, significant re-framing or different positioning of one’s research may not be seen as plagiarism.

[2] The Sorites Paradox (or Paradox of the Heap) is generally attributed to Eubulides of Miletus and states that it is impossible to establish the exact point of transition between two states along a continuum.

Even more complicated seem phenomena like cryptomnesia. In this case the distinction suggested between intentional and unintentional plagiarism plays a crucial role and assumes complex contours. Cryptomnesia refers to a recollection that is perceived as a new creation. In other words, as van den Berk states, “one remembers something without realizing it is a recollection” (2012: 3) and the process may regard one’s own words or somebody else’s. The phenomenon has long been investigated (see Jung’s approach to cryptomnesia; see inter alia Jung 1970[1905]) and is an example of the complexities underlying the process of defining and detecting plagiarism and self-plagiarism.

1.3. Concealment

A further distinction among different types of plagiarism regards their level of concealment, which can be distinguished between exact, near-exact and concealed plagiarism. This is not to say the first two categories do not imply an attempt to conceal plagiarism on the part of the author, but that there is a less significant presence of obfuscation processes (see Barrón-Cedeño *et al.* 2010: 2). In fact, all forms of plagiarism are based on some form of concealment. As Posner aptly remarks, “concealment is at the heart of plagiarism” (2007: 17), and if the reference to the other source is obvious and unmistakable, it will not constitute plagiarism. Therefore, for example, an allusion should not be seen as plagiarism “because the readership is expected to recognize the allusion” (Posner 2007: 18).

As has been pointed out, the concept of plagiarism is protean, and presenting a clear taxonomy is complex as different categories overlap and may be placed along a continuum. Figure 1 attempts to show some of the different categories of plagiarism:

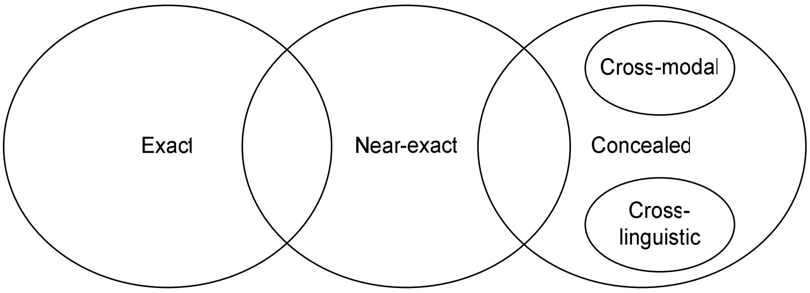

Figure 1. Types of plagiarism.

In its simplest form, exact plagiarism regards a verbatim replication of somebody else’s words without an appropriate quotation. Near-exact plagiarism may involve procedures such as modest changes in word order, the use of synonyms, rephrasing strategies, etc. Concealed plagiarism, instead, concerns the duplication of concepts rather than words, in the sense that the author does not reproduce portions of texts or expressions literally but reframes and reformulates certain concepts without citing the original source.

The first type of plagiarism is of course the easiest to detect, in that detection software products can uncomplicatedly detect exact replication of words that are reproduced in a certain sequence. Conversely, as we shall see, processes of obfuscation make automatic detection more complex. Two of the phenomena that may be included in the category of concealed plagiarism are cross-modal and cross-linguistic plagiarism. The former is based on representing the same concept by using a different modality, e.g. a picture instead of words. The latter is a very common process in academic writing. In this case the translation of a text into another language obviously implies that an exact representation of the same words is absent, but the same concepts are presented through a different tongue without acknowledging the original source.

2. Plagiarism detection

Plagiarism detection may be manual or computer-mediated. The two approaches are clearly interdependent and are generally combined in order to achieve reliable findings.

As regards automatic analyses, different types of plagiarism detection software have been developed in recent years. A wide variety of systems is now available and they are ascribable to two main methods: internal (also called intrinsic or endophoric) and external (or extrinsic or exophoric) detection. Figure 2 attempts to rudimentally show the main types of approaches to plagiarism detection.

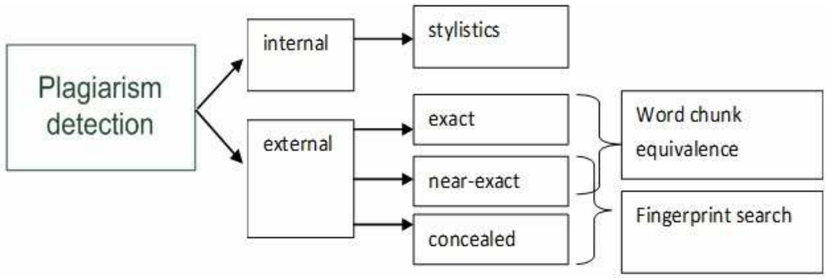

Figure 2. Main plagiarism detection approaches.

The principal internal detection method is based on stylistic analysis. This process aims at identifying chunks of a text which may not have been entirely written by the same author because they are characterized by a style that differs from the predominant one. This approach is often based on statistical evaluations focusing, for example, on lexical, syntactic and textual tendencies. The use of stylometry is now well consolidated and has been developing for several years (see Holmes 1998). For example, *Stylysis*[3] is a Web application based on a set of measures for stylistic analysis. More specifically, it is based on the computation of features such as Gunning Fog Index, sentence length, average word length, Honore's R function, Yule's K function, and Flesch-Kincaid readability test. The results show the sentences which differ from the general writing style and, thus, may be cases of plagiarism.

Similar procedures are based on the computation of other features. For instance, the parameters listed by Meyer zu Eissen *et al.* (2007: 361) are:

- text statistics;
- syntactic features;
- part of speech features;
- closed-class word sets to count special words;
- structural features.

In other words, intrinsic plagiarism detection is generally used without external reference corpora or collections as it is based on the identification of stylistic variations and discrepancies within a text. However, it should be noted that stylometric tools may also be used in external detection, more specifically as a "preprocessing step to an external plagiarism detection tool" (Stamatatos 2009: 38).

On a final note, internal detection should be combined with manual evaluation in order to achieve more reliable results. Indeed, it is clear that an author may display different stylistic features within the same text for reasons that are not related to plagiarism and therefore false positive results may occur.

Exophoric plagiarism is generally based on text comparison, checking the text under investigation against others (Barrón-Cedeño *et al.* 2010). Automatic systems based on external detection are often

[3] To learn more about Stylysis (developed by Barrón-Cedeño, Vallés-Balaguer and Rosso) and to access the stylistic analysis tool see: <http://memex2.dsic.upv.es:8080/StylisticAnalysis/en/index.jsp>

designed to detect not only sameness in word sequences but also concealed similarities which are identifiable through different procedures such as fingerprinting[4]. Several systems for the detection of plagiarism in natural languages are now available and they display a vast array of functions (for a comprehensive overview see Maurer *et al.* 2006). Some of the most widely used are:

- Turnitin by iParadigms
- WordCHECK by WordCHECKsystems
- Findsame by Digital integrity
- Eve2 by CaNexus
- CopyCatch by CFL Software Developments

These programmes are based on the comparison between the document being investigated and a collection of documents assumed to be original. Other online tools for plagiarism detection are available at:

- www.PlagAware.com
- www.PlagScan.com
- www.CheckForPlagiarism.net
- www.PlagiarismDetection.org

3. Using corpora

3.1. Corpora for plagiarism detection and research

Plagiarism detection and research may make use of different kinds of corpora. By and large, a corpus may be defined as "an electronically stored collection of samples of naturally occurring language" (Hunston 2006: 234). More specifically, McEnery and Wilson (2001:

[4] Fingerprints can be defined as sets of "integers created by hashing subsets of a document to represent its key content" (El Bachir Menai 2012: 841).

197) specify that the term ‘corpus’ can refer to: “(i) (loosely) any body of text; (ii) (most commonly) a body of machine-readable text; (iii) (more strictly) a finite collection of machine-readable text, sampled to be maximally representative of a language or variety”.

In this chapter I argue that corpora may be fruitfully employed in plagiarism detection and, in particular, in two main ways. First of all, in the case of external plagiarism detection, authentic corpora can be used as monitor collections against which to check the text under investigation. Secondly, corpora can be specifically created to include instances of plagiarism, with the aim of reaching advancements in the understanding of how to improve plagiarism research tools.

As regards the first approach, a text may be checked against collections of locally stored documents (as in the case of *PF*, which will be described in Section 3.2), general or specialized corpora, online databases, or the web. In the case of corpora, the choice of what kind of corpus should be used for plagiarism detection (monolingual/ multilingual, spoken/written, general/specialized, synchronic/diachronic, etc.) depends on the objectives and the resources available. For example, a corpus such as CADIS[5], Corpus of Academic English, compiled at the University of Bergamo, may be used for preliminary text comparisons. This corpus consists of academic texts written in English and is subdivided into four different disciplinary areas (i.e. applied linguistics, economics, law and medicine). According to the topic of the text under investigation, it is possible to limit the search to a certain discipline, therefore optimizing the search time.

In relation to other collections of texts, such as archives, databases or the web, it may be argued that they may not technically be seen as corpora. Indeed, they lack some of the defining characteristics of a corpus, such as representativeness, balance or size. However, as we shall see, for some specific purposes related to plagiarism detection, they may be used as corpora in the loose sense identified by McEnery and Wilson (2001: 197). For example, although not corpora in the traditional senses, databases provided by a university library system may also serve as monitor collections, and

[5] For an overview see <http://www.unibg.it/cerlis/cadis>.

the same holds true for the web (for a comprehensive discussion of the issues related to considering the web as a corpus see Kilgarriff/ Grefenstette 2003).

The second way corpora may be used to investigate plagiarism is by specifically creating them for detection purposes. The PAN Plagiarism Corpus[6] is probably the most representative example. It contains documents in which plagiarism has been inserted automatically and manually in order to allow for the evaluation and the assessment of automatic plagiarism detection algorithms (Potthast *et al.* 2010). The PAN corpus contains cases of artificial and simulated plagiarism and its development is based on the principle that the construction of training corpora of this kind "can be automated, and hence be done on a large scale" (Potthast *et al.* 2010: 997). This corpus, being artificially created for the evaluation of automatic plagiarism detection tools (Barrón-Cedeño *et al.* 2010), is widely used for research purposes. It predominantly includes cases of monolingual plagiarism but some cross-linguistic examples are also present.

Working along the same lines as PAN, another collection was compiled to focus specifically on cross-linguistic plagiarism, namely, ECLaPA (Europer Cross-Language Plagiarism Analysis). This is an artificially compiled test collection which includes instances of cross-language plagiarism detection (see Pereira *et al.* 2010).

3.2. PF as an external plagiarism detection tool

If an evaluator (or self-evaluator) is aiming at a preliminary form of plagiarism search, a programme purely based on the identification of word chunk similarity may easily be created. With this objective in mind, we developed *PF*[7] (*PlagiarismFinder*) which works simply on

[6] The latest version is PAN-PC-11, downloadable at http://www.uni-weimar.de/ cms/medien/webis/research/corpora/corpus-pan-pc-11.html#c58437

[7] The software is in its testing phase but will soon be available. For information: patrizia.anesa@unibg.it or giuseppe.parrinello@gmail.com.

text similarity detection. It was developed in C#[8] and the reference corpus may be composed of files in different formats[9]. It is primarily intended for external detection where the text is checked against a specific collection of texts or against corpora. There are no limits to the number of files which may be used as reference and the search can be carried out against specific files or entire folders or sub-folders.

Maurer *et al.* (2006: 1059) point out that there are three main approaches to detecting plagiarism:

- comparing a document to a body of texts (which may be stored locally);
- using a search engine;
- carrying out a stylistic analysis.

Following this categorization, I suggest here that the use of *PF* would fall within the first class. However, it can easily be combined with the other two approaches. Indeed, one can also check a part of text that is suspicious by using search engines, and external plagiarism detection can also be integrated with stylistic observations (Maurer *et al.* 2006: 1059).

PF lacks the benefits offered by more sophisticated programs (see Section 2) but may represent a fruitful tool in two main situations:

- to detect self-plagiarism in academic writing;
- to detect plagiarism in student assignments.

Detecting self-plagiarism may seem trivial in that authors should be fully aware of what they have previously published. However, cases of unintentional self-plagiarism are possible. For instance, when an author is working on several related publications, time pressure may lead to the unintentional duplication of sentences. Running a very

[8] C# is a modern, object-oriented programming language developed by Microsoft within the .NET initiative.

[9] While testing has been mainly based on txt files, processing can be done also using other formats, such as pdf files.

simple programme such as *PF* could prevent such unpleasant situations. Indeed, the text can effortlessly be checked against an author's collection stored locally, showing potential cases of duplication. Subsequent manual intervention can easily disambiguate the results.

The other main potential application of simple software such as *PF* lies in the educational field. It may certainly be argued that several institutions regularly require that their students submit their essays using a computerized system which automatically highlights potential cases of plagiarism. However, when such a system is not available, instructors can carry out a preliminary analysis by creating collections of students' assignments to be used as reference collections.

Locally stored collections are certainly homogenous regarding some of Sinclair's (2005) criteria to be considered in compiling a corpus, namely: mode, type, domain, language(s), location, and date. However, as has been mentioned, a collection of this type is not definable as a corpus in the strict sense (McEnery/Wilson 2001: 197). Therefore, in this case we are not talking about typical corpora but rather about *ad hoc* collections of texts.

Given its purposes, the creation of such collections does not generally require sampling. Moreover, the collections of texts which are used for plagiarism detection are not in line with the principle of representativeness described by Biber (1993: 243) as "the extent to which a sample includes the full range of variability in a population". It is clear that the texts collected for this specific purpose are not intended to fully depict a certain language or language variety. Indeed, in terms of size, an author's collection or a collection of students' assignments obviously cannot be representative of a language variety *in toto*. However, a collection of this type may include the total number of an individual's writings, and therefore it may even be seen as almost exhaustive in representing one's own variety. Similarly, a collection of student assignments may be considered as illustrative of a specific group's variety.

Representativeness should be also taken into account from a diachronic perspective in that "any corpus that is not regularly updated rapidly becomes unrepresentative" (Hunston 2002: 30). In line with the objectives of *PF*, the collection to be used as reference can simply be updated including new texts (both in the case of self-plagiarism and student plagiarism).

As with internal detection, external plagiarism detection also has some limitations. For instance, the recognition of an identical passage taken from another text could be based simply on a citation which has been correctly reported. However, as stated above, false positives can easily be double-checked manually. Moreover, plagiarism in its concealed form is realized through reformulations, paraphrases and other modification processes. Therefore, a detector based on word chunk identity such as *PF* can certainly serve as a preliminary tool, in that it is free, fast and easy to use. However, for more in-depth analyses it should be combined with more sophisticated tools that are also based, for instance, on fingerprint indexing (see Barrón-Cedeño *et al.* 2010). Another limitation of an approach based on chunk identity is that it does not solve the problem of multilingual plagiarism. However, new methods have been developed to identify instances that involve different languages (e.g. MLPlag, as illustrated in Ceska *et al.* 2008). Such methods would go beyond the scope of *PF* but may be used for more refined detection procedures.

Other issues are related to the fact that, as happens with any automatic detection system, plagiarism concerning material that is only available in a paper format is not detected. Moreover, as mentioned in Section 1, works produced by ghost writers or through a paper mill may also be difficult to identify as unauthentic if they do not display real cases of plagiarism.

On a final note, it should be underlined that *PF* does not discriminate between intentional and unintentional plagiarism, where the latter is caused by an inability to deal with quotations and references or by genuine mistakes which may occur in the process of rephrasing or paraphrasing. However, no software product seems to deal satisfactorily with the issue of intentionality. Therefore, an evaluator's opinion is inevitably necessary.

As mentioned above, the notion of plagiarism, like those of intellectual creativity and originality, assume different contours in different cultures and depend on individual factors. Moreover, the interpretation of plagiarism in an educational context is also based on the concept of teacher reflexivity, intended as the reflexive role of the teacher in the process of reading and interpreting a text (Sutherland-Smith 2008: 16), which may lead to different perceptions and interpretations of plagiarism.

It should also be noticed that the fight against plagiarism cannot be based solely on detection, but also on prevention. From a technical point of view, the use of passwords or encryptions in digital files can limit the ease of plagiarism practices. Secondly, as regards students, a clear policy focusing on the importance of originality as well as on the measures to be adopted in the case of plagiarism could be used as a deterrent. Indeed, as Emerson (2008: 193) aptly remarks, "the power of detection does not compensate for the breaking of the relationship between student and teacher" and detection should be accompanied by adequate prevention policies.

4. Conclusions

The overarching purpose of this chapter was twofold: to discuss potential systems to detect phenomena of plagiarism and self-plagiarism in writing and, more specifically, to observe how corpora could be used in this perspective. Firstly, I observed that different kinds of authentic corpora could be employed for text-external detection. Moreover, artificially created plagiarism corpora can be used in order to gain further knowledge in the area of automated plagiarism detection research and for the evaluation of current detection tools.

As regards plagiarism detection, I have argued for a multi-methodological approach which combines automatic and manual systems. In particular, in terms of automatic detection, this chapter highlights the necessity to check documents against self-plagiarism by using a locally stored collection of one's texts, in order to avoid possible cases of accidental self-plagiarism. I also observed the usefulness of attempting to detect other-plagiarism by using corpora or other collections not technically definable as corpora, such as archives or databases.

More specifically, I presented a simple detection program, *PF*, which may be used to check the document under investigation against a locally stored collection. When dealing with plagiarism detection, a web retrieval scenario is certainly more exhaustive, but, as has been mentioned, *PF* is merely intended as a preliminary step in the detection process. Moreover, if the aim is to avoid self-plagiarism, an author's collection can represent an exhaustive monitor collection. Similarly, a collection of assignments may be sufficient if the objective is to carry out a preliminary analysis of student assignments to prevent intra-class or inter-class plagiarism.

Ethical issues lie at the heart of plagiarism. Some scholars have argued against the use of detection software following the principle that it is unethical in that, essentially, it makes money from a crime. However, a simple tool like *PF* is completely free and therefore issues of this kind are minimized, and, despite its limitations, it represents a useful resource for preliminary detection.

To sum up, we can argue that no plagiarism detection program can be used in isolation. First of all, a good combination of automatic and manual practices is desirable. Secondly, in terms of automatic detection, different tools may be used according to the specific objectives and the resources available. Simple word chunk identity detection may be seen as a preliminary step towards the avoidance of self- or other-plagiarism in student assignments. However, when dealing with the education field, it should be noted that the attempt to limit plagiarism among students cannot be based solely on detection systems but should start with a good teaching approach that focuses on the consequences and the moral issues related to this practice. Similarly, in the case of self-plagiarism, automated systems cannot replace a careful and meticulous analysis of one's texts.

Further research in this field should also focus on qualitative analysis involving authors and students whose texts present some forms of plagiarism or self-plagiarism in order to gain a finer understanding of the reasons and the practices lying behind this phenomenon.

References

Anderson, Judy 1998. *Plagiarism, Copyright Violation and Other Thefts of Intellectual Property*. Jefferson: McFarland.

APA 2010. *The Publication Manual of the American Psychological Association*. 6th ed. Washington, D.C.: American Psychological Association.

Arwin, Christian / Tahaghoghi, Seyed M.M. 2006. Plagiarism Detection across Programming Languages. In Estivill-Castro, Vladimir / Dobbie, Gillian (eds) *Proceedings of the 29th Australasian Computer Science Conference* 48, Darlinghurst: Australian Computer Society, 277-286.

Barrón-Cedeño, Alberto / Potthast, Martin / Rosso, Paolo / Stein, Benno 2010. Corpus and Evaluation Measures for Automatic Plagiarism Detection. In Calzolari, Nicoletta / Choukri, Khalid / Maegaard, Bente / Mariani, Joseph / Odijk, Jan / Piperidis, Stelios / Rosner, Mike / Tapias, Daniel (eds) *Proceedings of the International Conference on Language Resources and Evaluation, LERC 2010*. Valletta: European Language Resources Association, 771-774. Retrieved from: http://users.dsic.upv.es/~prosso/resources/BarronEtAl_LREC10.pdf

Biber, Douglas 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4, 243-257.

Bloch, Joel 2012. *Plagiarism, Intellectual Property and the Teaching of L2 Writing*. Bristol: Multilingual Matters.

Buranen, Lise 1999. But I wasn’t Cheating: Plagiarism and Cross-cultural Mythology. In Buranen, Lise / Roy, Alice (eds) *Perspectives on Plagiarism and Intellectual Property in a Postmodern World*. NY: State University of New York.

Campbell, Cherry 1990. Writing with Others’ Words: Using Background Reading Text in Academic Compositions. In Kroll, Barbara (ed.) *Second Language Writing*. Cambridge: Cambridge University Press, 211-230.

Ceska, Zdenek / Toman, Michal / Jezek, Karel 2008. *Multilingual Plagiarism Detection*. Berlin: Springer-Verlag.

Clough, Paul / Gaizauskas, Robert 2009. Corpora and Text Re-use. In Lüdeling, Anke / Kytö, Merja (eds) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter. 1249-1271.

El Bachir Menai, Mohamed 2012. Detection of Plagiarism in Arabic Documents. *International Journal of Information Technology and Computer Science* 10, 80-89.

Emerson, Lisa 2008. Plagiarism, a Turnit Trial, and an Experience of Cultural Disorientation. In Eisner, Caroline / Vicinus, Martha (eds) *Originality, Imitation, and Plagiarism.* Ann Arbor: The University of Michigan Press, 183-194.

Holmes, David I. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13/3, 111-117.

Howard, Rebecca 2007. Understanding Internet Plagiarism. *Computers and Composition* 24/1, 3-15.

Hunston, Susan 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, Susan 2006. Corpus Linguistics. In Keith, Brown (ed.) *The Encyclopedia of Language and Linguistics*. Boston, MA: Elsevier, 234-248.

Jung, Carl Gustav 1970[1905]. Cryptomnesia. *Collected Works*, Vol. 1. London: Routledge and Kegan Paul, 95-106.

Kilgarriff, Adam / Grefenstette, Gregory 2003. Introduction to the special issue on the web as corpus. *Computational linguistics* 29/3, 333-347.

Maurer, Hermann / Kappe, Frank / Zaka, Bilal 2006. Plagiarism: A Survey. *Journal of Universal Computer Science* 12/8, 1050-1084.

McEnery Tony / Wilson Andrew 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Meyer zu Eissen, Sven / Stein, Benno / Kulig, Marion 2007. Plagiarism Detection without Reference Collections. In Decker, Reinhold / Lenz, Hans-Joachim (eds) *Advances in Data Analysis*. Berlin: Springer, 359-366.

Murray, Laura J. 2008. Plagiarism and Copyright Infringement: The Costs of Confusion. In Eisner, Caroline / Vicinus, Martha (eds) *Originality, Imitation, and Plagiarism.* Ann Arbor: The University of Michigan Press, 173-182.

Pecorari, Diane 2003. Good and Original: Plagiarism and Patchwriting in Academic Second-language Writing. *Journal of Second Language Writing* 12, 317-345.

Pereira, Rafael C. / Moreira, Viviane / Galante, Renata 2010. A New Approach for Cross-Language Plagiarism Analysis. In Agosti, Maristella / Ferro, Nicola / Peters, Carol / de Rijke, Maarten / Smeaton, Alan (eds) *Proceedings of the CLEF 2010 Conference on Multilingual and Multimodal Information Access Evaluation*. Padua: Springer, 15-26.

Posner, Richard 2007. *The Little Book of Plagiarism*. New York: Pantheon.

Potthast, Martin / Stein, Benno / Barrón-Cedeño, Alberto / Rosso, Paolo 2010. An Evaluation Framework for Plagiarism Detection. *23rd International Conference on Computational Linguistics (COLING 10)*. Retrieved from: http://www.uni-weimar.de/medien/webis/publications/papers/stein_2010p.pdf

Sinclair, John 2005. Corpus and Text - Basic Principles. In Wynne, Martin (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1-16.

Stamatatos, Efstathios 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *Proceedings of the 3rd International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse.* Retrieved from: http://ceur-ws.org/Vol-502/paper8.pdf.

Stepchyshyn, Vera / Nelson, Robert 2007. *Library Plagiarism Policies*. Chicago: American Library Association.

Sutherland-Smith, Wendy 2008. *Plagiarism, the Internet, and Student Learning: Improving Academic Integrity*. New York: Routledge.

van den Berk, Tjeu 2012. *Jung on Art: The Autonomy of the Creative Drive*. New York: Routledge.