Marina Bondi and Matteo Di Cristofaro

MoReThesisCorpus

Documenting Academic Language as Used in the Theses Submitted to the University of Modena and Reggio Emilia

Abstract

The article discusses the on-going process for the creation of the MoReThesisCorpus, outlining its major characteristics and offering an account of the considerations and issues involved so far. The corpus, composed of the theses submitted to the University of Modena and Reggio Emilia between 2011 and 2020, is being developed as part of the project CAP ('Comunicazione Accademica e Professionale;' Academic and Professional Communication), and is meant to foster research into academic language in a cross-disciplinary discourse perspective, as well as to facilitate the production of educational materials aimed at university students. It aims at supporting the acquisition of discipline-related vocabularies and styles to improve the learning of academic writing through corpus tools and resources, following a data-driven learning approach. Technical details surrounding the acquisition and subsequent processing of the data are discussed, along with considerations on a number of issues pertaining both to computer science and linguistics, directly impinging on the capability of the corpus to correctly support an investigation of academic discourse across different languages and disciplines.

Keywords: *corpus linguistics, EAP, ESP, Master's theses, PhD theses*

The spread of English across cultural and linguistic boundaries has been a major influence in the world of research and higher education, leading to an increasing use of different varieties of English worldwide in academic communication and to a growing literature on intercultural perspectives on research writing (see, inter alia, Mur-Dueñas and Šinkūnienė 2018). The global dimension of academic activity, publications and research has given increasing importance to the use of English in the international academic panorama, thus bringing about changes in the study and teaching of English for Academic Purposes (EAP).

The field is characterised by a recent interest in English-medium areas of discourse, such as English for research publication purposes and English-as-a-Medium-of-Instruction (EMI) in higher education. The increasing use of English in many professional and academic contexts has played a key role in expanding its teaching at many universities in the world, as well as determining a growth of teaching in English. Studies on research publication purposes, for example, highlight recent changes in the English used in publishing (Hyland and Jiang 2019) and the interrelation between authors' mother languages and the international publication system, while also showing how researchers develop genre awareness in their disciplinary fields (see Flowerdew and Habibie 2022; Corcoran, Englander and Muresan 2019). Studies on EMI higher education programmes in Europe, on the other hand, suggest that students are not only fully exposed to English during class, but are also actively engaging with English to prepare exams, presentations, essays and final dissertations, as they develop specific academic literacies. This area has long attracted attention to the distinctive features of EMI contexts in higher education (e.g., Campagna and Pulcini 2014; Dearden 2014; Costa and Coleman 2013; Fortanet-Gómez 2013; Doiz et al. 2012), with special reference to classroom interaction and lecturing styles (e.g., Aguilar-Pérez and Kahn 2022; Bier 2022; Costa and Mair 2022; Doiz and Lasagabaster 2022; Jensen and Thøgersen 2020).

Research has shown great interest in the analysis of the kind of English that is used, taught and studied in the various higher educational contexts as well as the different attitudes and policies adopted in different Anglophone contexts. In this perspective, it may be worth investigating how EAP is shaped by local and global linguistic forces, starting from the distinctiveness of the sociolinguistic contexts in which English is studied and the functional ranges and domains in which it is used. When aiming "to represent the cross-cultural and global contextualization of the English language in multiple voices" in the form of World Englishes (Kachru, Kachru and Nelson 2006, 1), it is important to say that English has become plural and Anglophone trends are open to a lot of variation in academic discourse. Divergences and convergences across different academic Englishes can be more easily studied now that large amounts of data in English are made available in electronic formats. When looking at scholars from different lingua-cultural backgrounds, it is easy to see for example variation in aspects of authorial identity, stance and audience engagement (e.g., Mur-Dueñas and Šinkūnienė 2016).

The dominance of English in research has also created the idea of a discrimination between native and non-native speakers, as scholars from all over the world initially lived this need to be operational in English with "anxiety" and as a policy-imposed necessity (Ferguson, Pérez-Llantada and Plo 2011; Pérez-Llantada et al. 2011; Lillis and Curry, 2010). From other points of view, however, if English is perceived as a language for communication and not for identification, non-native speakers can also be shown to contribute to the richness of perspectives in their disciplinary communities (see also Pérez-Llantada, 2014; Flowerdew 2001; 1999; Suresh Canagarajah 1996) and to provide researchers with global access, greater visibility and possibilities for international collaboration. From the point of view of language studies, attitudes have moved from looking at these varieties of academic English as "defective forms of English" (Greenbaum 1996, 17) to studying them as natural developments of the widespread use of English in academia. Academic Englishes can be seen as a parallel phenomenon to that of World Englishes (Mauranen 2018) in a context of growing acceptance of the plurilingual and multicultural diversity of scholarly communication (Sano 2002, 49): non-canonical but comprehensible usage has been shown to be accepted for publication in journals (Hynninen and Kuteeva 2017; Rozycki and Johnson 2013) and to be operational in educational contexts (Mauranen 2012).

The major role of non-native speakers of English has drawn increasing attention to the approach of English as a Lingua Franca (ELF), the use of English as shaped by its users. Academic English needs to be dissociated from its native lingua-cultural roots and its Anglo-American English (core) variety (Mauranen, Pérez- Llantada and Swales 2020). It can be seen as a Lingua Franca, with scholars using English in their own diverse ways, irrespective of standardization and national boundaries. In contexts where mutual understanding and explicitness can be more important than correctness (Mauranen 2012; Mauranen, Hynninen and Ranta 2010), communicative activity often involves endonormative elements (Hynninen 2016). With EMI instruction, education is also experiencing increasing situations of intercultural use of English in academic contexts. This often means striving for greater explicitness in search for mutual understanding (Mauranen, Hynninen and Ranta 2010, 184), for example by means of indicating local organization, negotiating topics and careful use of metadiscourse (references to the discourse itself, contributing to the organization of the evolving text rather than to the subject matter) (Mauranen 2012). Communicative efficiency is shown to play a central role in determining the choice of structures that enhance clarity and explicitness without ignoring the news of conciseness (Wu, Mauranen and Lei 2020).

Debate over English in academic contexts has also acknowledged that academic writing is not part of a "natural acquisition" of a language, but it is acquired through lengthy formal education (Ferguson, Pérez-Lantada and Plo 2011). Features that are specific to a language variety—such as technical taxonomies, lexical density, rhetorical structures and grammatical patterns—have to be learnt by native and non-native speakers alike (Tribble 2017; Römer and Arbor 2009). The use of appropriate forms and rhetorical structures that suit specific academic contexts involves an effort from the native and the non-native speakers, both in the mother tongue and in an additional language. This leads to a process of a constant remodelling of the structures used to communicate, whether one is acquiring academic competence in one's own first language or not. This new perception of the needs of both native and non-native speakers to familiarise with and master the conventions of academic discourse suggests that EAP studies would profit from closer study of the early stages of acquisition of these practices. An area that certainly needs to be studied further is that of MA/MSc and PhD thesis writing. Dissertations can help trace important stages in the development of academic literacy through the formative years, when students—as novices—experiment with developing their identity as academic writers, to doctoral education, when they actually start shaping their disciplinary voices as rhetorical subjects (see Paré, Starke-Meyering and McAlpine 2011). Studying the rhetorical and linguistic features of dissertations from Master's level to doctoral level can help understand developmental perspectives, as well as disciplinary variation and variation across first and additional languages.

Focusing on a specific context (that of Italian Universities), we chose the University of Modena and Reggio Emilia as a case study that offers a wider range of disciplines and educational contexts both with Italian and with English as a medium of instruction. We are thus developing a corpus of dissertations from the official repository of the University and hope to make it available in a modular structure that could potentially allow for comparison across different stages of education and across different languages.

The need for a corpus of dissertations has certainly been acknowledged by others. Corpora that are available for online or onsite consultation include for example the University of Helsinki's English/ French/ German/ Russian/ Spanish/ Swedish E-thesis Corpus (available for consultation on KORP)[1] or the Reading Academic Text Corpus (available on site).[2] These corpora, however, do not allow for a specific case study of the English written by students of Italian Universities, or any cross-cultural comparison of English and Italian. For academic Italian we do have the Athenaeum Corpus,[3] which centres on the University of Turin, but includes other forms of academic Italian (the university magazine, mails, circulars etc.) and excludes dissertations. The MoReThesis corpus that we are developing would fill an important gap in allowing researchers to study the language of theses written at Italian universities in different languages (along the lines of the UH corpus), with a special focus on Italian and

[1] The corpus is available from https://www.kielipankki.fi/corpora/e-thesis/. All websites last visited on 21/09/2022.

[2] https://www.reading.ac.uk/internal/appling/corpus.htm.

[3] http://www.bmanuel.org/projects/at-HOME.html.

English.

The present article discusses the on-going process for the creation of the MoReThesisCorpus, outlining its major characteristics and offering an account of the considerations and issues involved so far. The corpus, composed of the theses submitted to the university of Modena and Reggio Emilia between 2011 and 2020, is being developed as part of the project CAP ('Comunicazione Accademica e Professionale'; Academic and Professional Communication), and is meant to foster research into academic discourse in a cross-cultural and cross-disciplinary perspective (see e.g. contributions in Charles, Pecorari and Hunston 2009; and in Hyland and Bondi 2006), as well as to focus on second language writing in academic settings and facilitate the production of educational materials aimed at university students. It aims at supporting the acquisition of discipline-related rhetorical structures and lexico-grammatical resources to improve the learning of academic writing (see Tribble 2002) through corpus tools and resources (see Flowerdew 2015; 2009), following a data-driven learning approach (Chambers 2012).

The rest of the article is organised as follows: Section 1 presents the procedures used to develop the MoReThesisCorpus starting from the University repository; Section 2 analyses the problems encountered and the decisions taken; Section 3 briefly exemplifies the potential use of the corpus for a study of its modules in intralingual or cross-linguistic perspectives; Section 4 draws some conclusions and points at directions for future research.

1. Project workflow: from MoReThesis to MoReThesisCorpus

The MoReThesisCorpus contains the theses and dissertations submitted by the students of Università degli Studi di Modena e Reggio Emilia for their BA, MA, and PhD degrees, as extracted from the public PDF files stored on the MoReThesis digital repository that archives materials from 2011 onwards into a browsable catalogue.

Identified language[4]	**N. of theses**	**Tokens**
English	1,063	24,580,351
French	7	139,773
German	8	162,722
Italian	2,772	60,825,356
Spanish	16	556,521
unidentified	2	21,315
TOTAL	3,868	86,286,038

Tab. 1: Languages and tokens included in the MoReThesisCorpus

[4] See Section 'Issues and Coding Decisions' for details concerning the identification of languages.

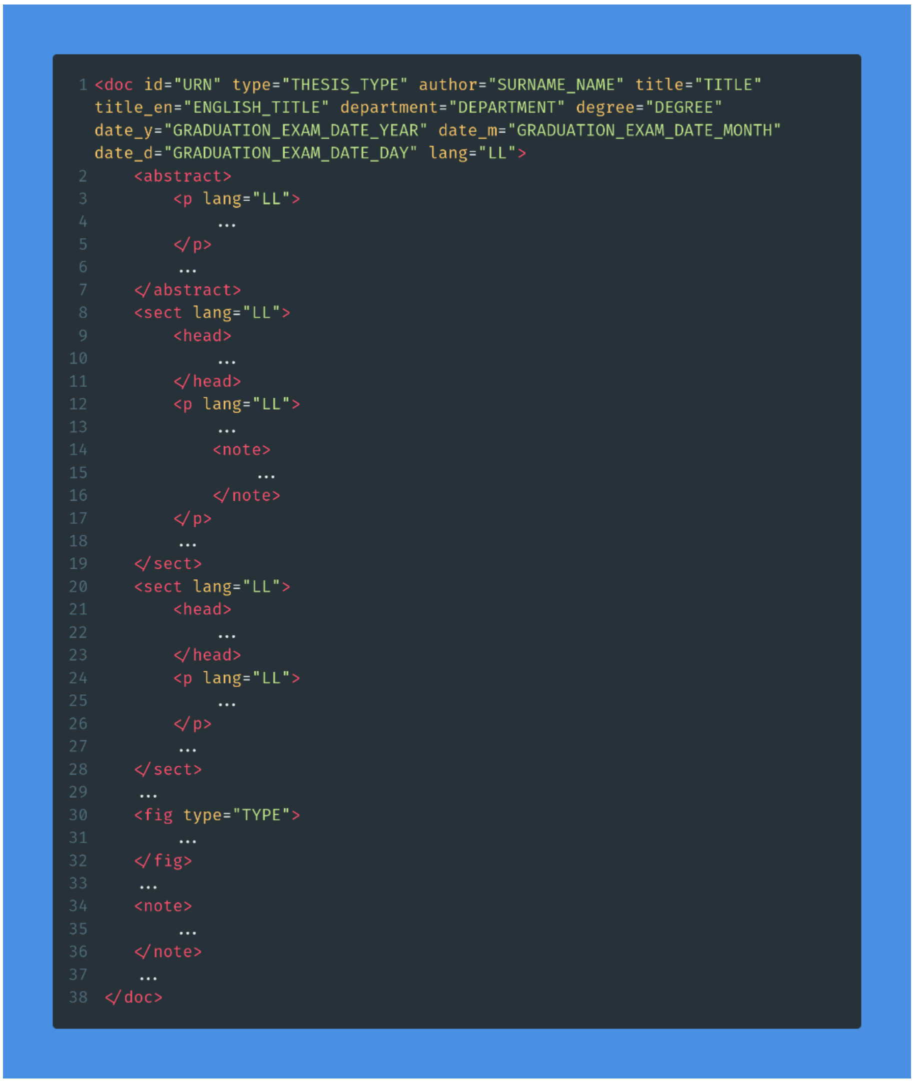
Corpus creation was started in February 2021, and as such it only took into account theses submitted until 2020, for a total of approximately 4,000 theses amounting to over 85 million tokens. Table 1 reports the number of theses available for each one of the identified languages (more details on the identification of the languages are discussed in Considerations) and their respective total number of tokens.

The basic structure of the MoReThesisCorpus revolves around documents, where each document represents a single thesis formatted into XML, following the "Modest XML" proposal by Hardie (2014) allowing the corpus to be enriched with metadata for filtering and querying procedures through corpus tools such as CQPweb (Hardie 2012) and SketchEngine (Kilgarriff et al. 2004; 2014). The adoption of the "Modest XML" proposal draws from Hardie's realisation that existing standards (e.g., TEI) are often excessive for the majority of uses a corpus may have, and rely on a complex system of tags that require extended efforts to be correctly applied. The tiniest mistake may consequently render the corpus unusable, and software tools may struggle with the amount of metadata required to load. For these reasons, 'Modest XML' makes full use of the XML format flexibility and standard, allowing users to create a limited amount of ad-hoc tags centered around the researcher's needs, while taking advantage of the full power of the format. Furthermore, being fully XML-compliant, the custom structure can easily be converted into e.g. TEI standard through the use of XML libraries such as lxml.

Collection, extraction, and formatting of the data was conducted through the use of custom Python scripts,[5] producing the corpus structure exemplified in Fig. 1: the following sections provide a description of the structure and of the scripts used to build the corpus. The contents of each thesis are assigned a root tag element named <doc>, containing a number of attributes describing metadata details created during the data collection and formatting processes. The textual contents of each thesis are structured into four possible elements, identified by the tag elements <abstract>, <sect>, <fig>, and <note>, respectively containing the abstract, the actual text of the thesis, the captions of the figures, and the (foot)notes. The additional structural tag element <p> separates the paragraphs, containing the textual data of the thesis in tokenised, POS tagged, and lemmatised format (using Stanza; see point 5. 'Creation of the final corpus' below for more details). In order to describe the procedures adopted and the value each metadata element has, let us start by defining how these are stored onto the MoReThesis platform. Each thesis is assigned a catalogue card, identified by a Unique Resource Number (URN) in the format exemplified below (where N is a single digit).

[5] The scripts are available at https://github.com/mdic/morethesiscorpus.

- etd-NNNNNNNN-NNNNNN



```
<doc id="URN" type="THESIS_TYPE" author="SURNAME_NAME" title="TITLE"
title_en="ENGLISH_TITLE" department="DEPARTMENT" degree="DEGREE"
date_y="GRADUATION_EXAM_DATE_YEAR" date_m="GRADUATION_EXAM_DATE_MONTH"
date_d="GRADUATION_EXAM_DATE_DAY" lang="LL">
    <abstract>
        <p lang="LL">
            ...
        </p>
        ...
    </abstract>
    <sect lang="LL">
        <head>
            ...
        </head>
        <p lang="LL">
            ...
            <note>
                ...
            </note>
        </p>
        ...
    </sect>
    <sect lang="LL">
        <head>
            ...
        </head>
        <p lang="LL">
            ...
        </p>
        ...
    </sect>
    ...
    <fig type="TYPE">
        ...
    </fig>
    ...
    <note>
        ...
    </note>
    ...
</doc>
```

Fig. 1: Corpus structure

The URN therefore serves as a unique identifier for both the details concerning the thesis—as explained further below—and as a constructor for the link to each thesis' catalogue card, as exemplified in the link reported below.

- https://morethesis.unimore.it/theses/available/etd-01162017-111702/

Besides the link(s) to the thesis' file(s), each catalogue card provides an HTML table with metadata data-points. These are reported in table 2 below, where for each metadata point its original Italian label is presented alongside its English translation. Two additional columns report whether the metadata point was preserved as part of the corpus metadata, and a brief description of what it describes.

Original field name	English translation	In corpus	Description
Tipo di tesi	Thesis type	X	Whether it is an MA or PhD thesis
Autore	Author	X	Name of the author, in the format SURNAME, NAME
URN	URN	X	Unique identifier of the thesis
Titolo	Title	X	The original title of the thesis
Titolo in inglese	Titole in English	X	The English title of the thesis (required by some departments and degrees)
Struttura	Department	X	The name of the department under which the student has graduated
Corso di studi	Degree	X	The name of the degree under which the student has graduated
Commissione	Members of Degree Committee		The list of members that were part of the degree committee; two additional fields are included in the sub-table, ‘Nome Commissario’ (Name of the member) and ’Qualifica’ (Role)
Parole chiave	Keywords	X	A list of keywords related to the thesis, as provided by the student
Data inizio appello	Graduation exam date	X	The date on which the graduation took place, in the form YYYY-MM-DD
Disponibilità	Availability	X	A caption indicating the status of the thesis, chosen among a list of predefined items and indicating whether the PDF file is publicly available
Riassunto analitico	Abstract	X	The abstract of the thesis, in Italian and English (the latter is optional)
File	File	X	The link to the PDF file, along with its size, plus a link to the form for contacting the author of the thesis

Tab. 2: Metadata available on the MoReThesis platform for each thesis

In addition to the metadata contained in each thesis' catalogue card, the MoReThesis platform is structured so as to categorise each thesis using one of eight possible labels (L1, LS, D1, D2, LC6, LM5, LM6, LM),[6] indicating the type of degree it was submitted for. These labels are also employed for the creation of HTML pages containing direct links to theses that belong to each category. For example, the link reported below points to the page listing all the theses (and relative hyperlinks) categorised as type D2 (i.e., PhDs).

- https://morethesis.unimore.it/theses/browse/by_tipo/D2.html

The creation of the corpus started from the categories' HTML pages, and was achieved through a set of Python scripts[7] that operated the following procedures:

1. acquisition of the HTML catalogue cards (i.e., the web pages hosted on the MoReThesis platform); script S01
2. extraction of the metadata points from the HTML tables, saved to a spreadsheet; script S02
3. acquisition of the PDF files (when publicly available); script S03
4. extraction of the text and additional markup from the PDF files; script S04
5. creation of the final XML corpus files by merging the extracted text and the extracted metadata (collected in step 2); script S05

Each procedure is presented separately, documenting the main steps and highlighting a number of considerations and issues faced during the data processing. The operations were split across multiple scripts to allow researchers an evaluation of the output data at different stages of the data processing, ensuring that potential errors in the final corpus are caught as early as possible.

[6] L1: laurea vecchio ordinamento (Master's degree, single cycle of 4 years); LS: laurea specialistica (Master's degree, second cycle, after a B.A.); D1: dottorati di ricerca (PhD); D2: dottorati di ricerca "riformati" (PhD); LC6: laurea specialistica a ciclo unico (single-cycle Master's degree); LM5, LM6: laurea magistrale a ciclo unico (single-cycle master's degree); LM: laurea magistrale (Master's degree).
See https://wiki.u-gov.it/confluence/display/ESSE3/Normativa+e+Tipo+Corso+di+Studio.

[7] Scripts are available at https://github.com/mdic/morethesiscorpus.

1.1 Acquisition of the HTML catalogue cards

The eight HTML category pages containing the links to the catalogue cards for the theses belonging to each category were manually downloaded and used as input for the first script (S01), which identifies and extracts the link to every thesis catalogue card (i.e., its catalogue web page) included in the category, such as the one reported in 3. Catalogue cards in HTML format downloaded to a local PC were then used as input for script S02.

1.2 Metadata extraction

Using the links to each thesis' web page, script S02 collects the HTML catalogue cards to a local folder. These are 1:1 copies of the pages hosted on the MoReThesis platform, and they contain the metadata HTML table previously discussed, as well as direct links to the PDF file(s) (see next section for more details). The metadata points included in each catalogue card is reported in table 1: these were extracted by parsing the HTML table, and saved to a spreadsheet where each row represents a thesis. These metadata points are later used in script S04 as source for the corpus metadata.

1.3 Acquisition of the PDF files

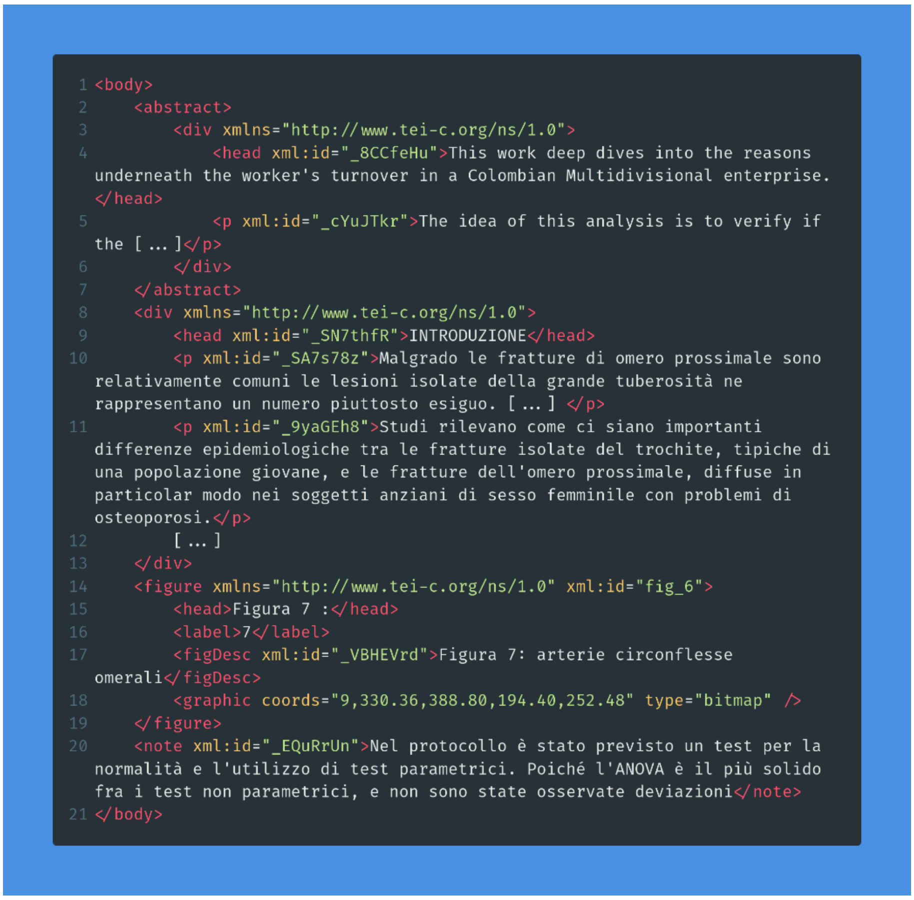
As previously mentioned, only public files were employed for the creation of the corpus: each student may in fact choose—when uploading their thesis onto MoReThesis—three levels of accessibility to the PDF file(s): publicly available, private (available on request), or publicly available after a number of months from the discussion (what is commonly labelled as 'embargo'). Additionally, no standard exists as to how the PDF files should be structured; hence a thesis may be contained in one single file, or split across multiple ones (e.g., each one containing a different chapter). Script S03 downloads every publicly available file and renames it according to the syntax exemplified below, where URN is the Unique Resource Number, N a progressive number used to identify in which position (from top to bottom) the file appears in the catalogue card, and FILENAME is the original filename assigned by the author of the thesis.

- URN_N-FILENAME.pdf

The downloaded PDF files are then used as source for the next script (S04).

1.4 Text and markup extraction

Using the metadata spreadsheet created through script S02 and the downloaded PDF files, script S04 employs the machine-learning Python tool GROBID[8] to identify the document layout and extract both the textual contents and the document structure in XML markup. The tool—originally developed for "extracting information from scholarly documents"[9]—provides the ability to automatically identify and extract the sections (e.g., abstract, chapters, references, figures, notes) of a document, converting them into XML-TEI format. The XML-TEI header introduced by GROBID was discarded during data processing (see description below), and only the contents of the thesis and their relevant tag elements were preserved; these are structured inside of a main <body> tag element according to the meta-structure exemplified in Fig. 2 (only those elements utilised for the construction of the corpus are exemplified).



```
<body>
    <abstract>
        <div xmlns="http://www.tei-c.org/ns/1.0">
            <head xml:id="_8CCfeHu">This work deep dives into the reasons
underneath the worker's turnover in a Colombian Multidivisional enterprise.
</head>
            <p xml:id="_cYuJTkr">The idea of this analysis is to verify if
the [ ... ]</p>
        </div>
    </abstract>
    <div xmlns="http://www.tei-c.org/ns/1.0">
        <head xml:id="_SN7thfR">INTRODUZIONE</head>
        <p xml:id="_SA7s78z">Malgrado le fratture di omero prossimale sono
relativamente comuni le lesioni isolate della grande tuberosità ne
rappresentano un numero piuttosto esiguo. [ ... ] </p>
        <p xml:id="_9yaGEh8">Studi rilevano come ci siano importanti
differenze epidemiologiche tra le fratture isolate del trochite, tipiche di
una popolazione giovane, e le fratture dell'omero prossimale, diffuse in
particolar modo nei soggetti anziani di sesso femminile con problemi di
osteoporosi.</p>
        [ ... ]
    </div>
    <figure xmlns="http://www.tei-c.org/ns/1.0" xml:id="fig_6">
        <head>Figura 7 :</head>
        <label>7</label>
        <figDesc xml:id="_VBHEVrd">Figura 7: arterie circonflesse
omerali</figDesc>
        <graphic coords="9,330.36,388.80,194.40,252.48" type="bitmap" />
    </figure>
    <note xml:id="_EQuRrUn">Nel protocollo è stato previsto un test per la
normalità e l'utilizzo di test parametrici. Poiché l'ANOVA è il più solido
fra i test non parametrici, e non sono state osservate deviazioni</note>
</body>
```

Fig. 2: Meta-structure with tag elements

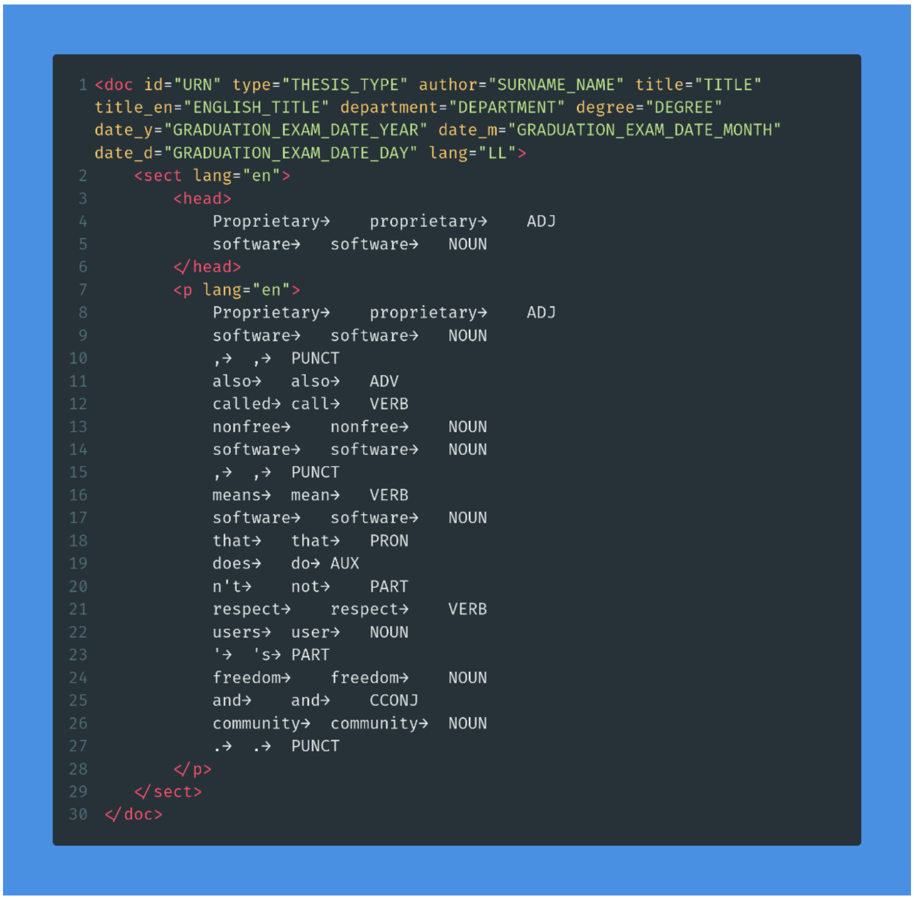
Once converted into XML-TEI format, the files are used as input to access the thesis' contents, while preserving the five structural elements (abstract, section, figures, notes and paragraphs) previously discussed in Fig.1; for the figures, only their description—included by GROBID in

[8] https://github.com/kermitt2/grobid.
[9] https://github.com/kermitt2/grobid.

the tag element <figDesc>—was preserved inside the <fig> tag element (see Fig. 1). During the development of the script for this procedure, a number of considerations were taken into account by the team, due to two potential sources of issues: i) the structure of the PDF files; ii) the language used for the thesis. GROBID is only able to correctly identify e.g., the abstract section if the document follows a specific structure, and this is not always the case in the PDF files under scrutiny. This first issue is connected to the second one, as not all the theses are written in English, meaning therefore that the section containing the abstract may not always be labelled as "Abstract"—further language-related issues are documented in 'Considerations.' The textual contents were processed through script S05 that only reads a limited set of TEI tag elements (the ones exemplified in 6) and extracts the textual content of each one. Additional metadata extracted through script S02 were added during content extraction and inclusion in the relevant tag elements. At this stage language identification was applied to each paragraph (more details in 'Considerations') and one XML file was created for each thesis, using the URN as filename; this step ensures that, in cases where more than one PDF file was submitted by the student for their thesis, all the relevant textual data are contained into a single file, with contents organized according to the order in which the original PDF files are shown on MoReThesis.

1.5 Creation of the final corpus

Each XML file from the previous step was then processed through script S05 that employs Stanza (Qi et al. 2020) for tokenisation, POS tagging, and lemmatisation, applied on a per-paragraph base. During this step Stanza applies the correct language model by reading the lang attribute of each paragraph tag created during the previous processing step. The annotation is included using the VeRticalized Text format (.vrt), a "token-oriented columnar text format" (Kielipankki 2021) whose structure resembles that of XML – and hence allows for the inclusion of metadata – with the difference that annotation is arranged so that e.g., POS and lemma details are horizontally arranged next to each token, separated by tabs. Vrt format is the standard input format for all corpus tools based on the IMS Open Corpus Workbench (CWB; Evert and Hardie 2011; Christ 1994) such as CQPweb and is supported by a number of other tools - including SketchEngine. Verticalized format is exemplified in Fig. 3, where the annotated verticalized text is included inside an empty corpus structure, using sample textual data; for clarity, the tab character is visually represented with the symbol →.



```
<doc id="URN" type="THESIS_TYPE" author="SURNAME_NAME" title="TITLE"
title_en="ENGLISH_TITLE" department="DEPARTMENT" degree="DEGREE"
date_y="GRADUATION_EXAM_DATE_YEAR" date_m="GRADUATION_EXAM_DATE_MONTH"
date_d="GRADUATION_EXAM_DATE_DAY" lang="LL">
    <sect lang="en">
        <head>
            Proprietary→    proprietary→    ADJ
            software→   software→   NOUN
        </head>
        <p lang="en">
            Proprietary→    proprietary→    ADJ
            software→   software→   NOUN
            ,→  ,→  PUNCT
            also→   also→   ADV
            called→ call→   VERB
            nonfree→    nonfree→    NOUN
            software→   software→   NOUN
            ,→  ,→  PUNCT
            means→  mean→   VERB
            software→   software→   NOUN
            that→   that→   PRON
            does→   do→ AUX
            n't→    not→    PART
            respect→    respect→    VERB
            users→  user→   NOUN
            '→  's→ PART
            freedom→    freedom→    NOUN
            and→    and→    CCONJ
            community→  community→  NOUN
            .→  .→  PUNCT
        </p>
    </sect>
</doc>
```

Fig. 3: Verticalized format

2. Issues and coding decisions

A number of decisions were taken during data collection and processing, following the continuous analysis conducted by the members of the team on the set of files produced by each one of the applied scripts. Among these, the ones concerning the presence of more than one language in a thesis arguably represent the major issue faced during the creation of the corpus. The majority of the departments at UniMoRe require students to write the abstract of the thesis in English, regardless of the language used in the thesis; and a number of departments may require a thesis to be submitted in English, or a language other than Italian. Students in language degrees may in fact e.g., submit a thesis written in German and with an English abstract. As such, both the main language of the thesis and the language used for the abstract produce—for the processing of the data—a number of combinations that can pose several issues when textual annotation is to be included. Paramount was therefore the identification of the language(s) used in each thesis prior to the annotation of the textual data, to ensure that the correct language model was applied and that no incorrectly annotated elements were introduced in the final corpus.

Consequently, during the extraction procedure (script S04) the textual data were processed through Lingua,[10] an open source module (the Python version was used in the script) developed for the identification of the language employed in portions of texts. In order to account for potential language-switching across subsequent sections, language identification was conducted on a per-paragraph basis, employing the paragraph sections (enclosed in <p> tag elements) identified by GROBID. Each paragraph was consequently run through Lingua, and the identified language was added to the attribute lang of each paragraph tag element, in the form of a two-letter label ('en,' 'it,' 'es,' 'fr,' and 'de' for English, Italian, Spanish, French, and German respectively; only these languages were included in the identification process as they are the only ones accepted by UniMoRe for the submission of a thesis); when no language was recognised (e.g., the paragraph only contains numbers, or a word that exhibits the same spelling in more than one language) the value 'none' assigned by Lingua was added as value of the lang attribute. The most frequently identified language in the paragraphs included in the section—as recognised by GROBID—was then assigned as value to the lang attribute of each <sect> tag element. A similar procedure was also adopted for determining the language of the thesis—as included in the attribute lang of the <doc> tag element—which was calculated by considering the most frequently identified language across all different <sect> tag elements. Beside the languages reported in Table 1, a number of theses (n=135) were not extracted as the PDF files contain images instead of selectable textual content. During the first stage of creation, these have therefore been discarded from the corpus and will require OCR procedures to extract textual data from images to be included in a future version of the corpus (see 'Conclusion and future directions').

While the lang attribute of each paragraph served during the corpus creation procedures to instruct Stanza, the remaining attributes in the <doc> and <sect> elements are included in the corpus to allow for filtering queries through corpus tools, and for the creation of purpose-specific sub-corpora.

As such the MoReThesisCorpus can be seen as a corpus-repository, allowing for the creation of ad-hoc sub-corpora tailored to a wide range of analytic purposes. These include: evaluation of the impact of English-as-a-Medium-of-Instruction (EMI); studies in English for Academic Purposes (EAP); creation of teaching materials based on real examples of language in use; investigations into English for Specific Purposes (ESP). The next section provides an overview of one such possible sub-corpora, composed of the PhD theses published on MoReThesis.

[10] https://github.com/pemistahl/lingua-py.

3. Application examples

Through the use of the metadata datapoint "Tipo di tesi" (Thesis type, see Table 2), it is possible to build a corpus documenting the language employed in PhD theses, composed of 635 complete theses. Details for these theses are reported in Table 3, where the total number of theses available for each degree is included along with the number of theses written in English and Italian (and their respective tokens), together with the number of theses written in languages other than English or Italian.

Degree	N. of theses	Eng. theses	Tokens (Eng.)	Ita. theses	Tokens (Ita.)	Other lang. theses
AGRI-FOOD SCIENCES, TECHNOLOGIES AND BIOTECHNOLOGY	48	48	1,403,661	0	0	0
CLINICAL AND EXPERIMENTAL MEDICINE	63	62	1,183,780	0	0	1
EARTH SYSTEM SCIENCES: ENVIRONMENT, RESOURCES AND CULTURAL HERITAGE	20	14	409,883	4	142,895	2
HEALTH SCIENCES AND TECHNOLOGIES	26	17	451,078	8	176,615	1
HIGH MECHANICS AND AUTOMOTIVE DESIGN AND TECHNOLOGY	62	38	1,112,994	20	610,372	4
HUMANITIES	47	9	841,908	36	3,088,355	2
INDUSTRIAL AND ENVIRONMENTAL ENGINEERING	20	12	292,957	8	134,973	0
INDUSTRIAL INNOVATION ENGINEERING	40	36	1,092,237	4	115,356	0
INFORMATION AND COMMUNICATION TECHNOLOGIES (ICT)	67	66	1,938,302	0	0	1
LABOUR RELATIONS	48	3	154,030	44	2,205,102	1
LEGAL SCIENCES	40	0	0	39	2,234,574	1
MODELS AND METHODS FOR MATERIAL AND ENVIRONMENTAL SCIENCES	37	36	916,786	0	0	1
MOLECULAR AND REGENERATIVE MEDICINE	43	43	786,136	0	0	0
NEUROSCIENCES	16	14	277,427	2	27,569	0
PHYSICS AND NANO SCIENCES	45	43	1,424,764	1	99,309	1
WORK, DEVELOPMENT AND INNOVATION	13	6	171,773	6	307,378	0

Tab. 3: Details of the PhD theses included in the MoReThesisCorpus

The corpus shows great potential for developing descriptions and teaching/learning materials for general EAP courses at doctoral level. If the whole set of data can help trace elements of general academic language, more specific descriptions can be obtained for specific areas.

When selecting for example all the PhD theses written in English, it would be possible to compare and highlight convergences and divergences across the main fields of knowledge, for example SH (Social Sciences and Humanities), LS (Life Sciences) and PE (Mathematics, physical sciences, information and communication, engineering, universe and earth sciences). Alternatively, one could focus on only one of these areas: the social sciences and humanities, for example, could be represented by PhD programmes such as those in the Humanities, Labour Relations, Legal Sciences and Work, Development, Innovation. The data shows that PhD dissertations are not often written in English in this area at UniMoRe, but it is still possible to use both the materials in English and those in Italian for a cross-cultural study. In other sectors, on the other hand, in the LS or PE sector, the use of English is largely dominant and the number of dissertations available allows for in-depth exploration of the use of English in thesis writing. The medical field, in particular, offers a wide range of examples from different subfields, well distributed over the years, so as to allow for an exploration of variation and change in the data from many different points of view: rhetorical structure, lexico-grammatical choices, pragmatic features etc. The identification of sections within the thesis also facilitates analysis (or practice) focused on specific sections, such as the literature review or the methodology. The data can obviously be used both for different descriptive and applied perspectives: ELF, academic Englishes, second language writing, literacy, curriculum design, materials design etc.

4. Conclusion and future directions

The MoReThesisCorpus is the first step towards the creation of an initiative aimed at producing a corpus-linguistics-ready version of the theses submitted each year to the university of Modena and Reggio Emilia. As such it represents an initial laboratory for the definition of replicable procedures meant to be automatically implemented and run every year. Work conducted so far has also highlighted the need to approach a number of issues—such as the previously mentioned presence of multiple languages and the absence of a unified structure for the submission of PDF files—that require a collaborative effort on behalf of multiple departments as well as administrative offices. Given the potential research and educational relevance of digital datasets of academic language, MoReThesisCorpus represents a tentative implementation of an automated solution through which the university can provide a yearly language corpus of academic language, fostering synchronic and diachronic inquiries into both academic language and students' proficiency.

The publication of such datasets also provides further opportunities for the development of deep learning approaches to academic discourse, such as the ones proposed by Becker et al. (2020). To further promote the value of the data, the corpus so far described is also being currently employed for the creation of an enriched version, containing a layer of manually annotated ad-hoc metadata describing the rhetorical features and the communicative functions of language—the work is being conducted through the use of INCEpTION (Boullosa et al. 2018; de Castilho et al. 2018; Klie 2018; Klie et al. 2018). This second version will also include a number of theses requiring OCR procedures for the extraction of the textual contents.

Parallel to the MoReThesisCorpus, a smaller corpus (MoReAbstractCorpus) composed of abstracts only is under development, to collect the abstracts submitted as part of the catalogue cards and included in the final submitted theses. During the collection of the data it was in fact noticed that the abstract included during the thesis submission process often differs from the abstract included inside the actual PDF file; this is due to how the submission process works, whereby students are asked to provide details about their thesis prior to the actual submission of the final PDF file. As such, two—often substantially different—versions of the abstract may be found; given the relevance that abstracts play in the study of academic writing (i.a. Bondi 2004, 2014), what might appear to be an issue for the creation of a corpus may in fact represent an invaluable source of linguistic data, documenting revisions and changes to the language used in abstracts, while providing further insights into academic writing in English and in Italian (see Flowerdew 2022 for an overview of the contribution of corpora to the teaching/learning of writing).

The MoReAbstractCorpus is just an example of the types of data and the type analysis that could be carried out on specific sections of the theses, whether in an intralinguistic or in a cross-linguistic perspective, to account for the features of academic discourse within Italian Universities. Another small project underway involves specific cases of courses where EMI is implemented; these could offer interesting materials for a focused study of the impact of EMI on thesis writing, for example by comparing theses written in English in different parallel courses in the same department or theses written in English or Italian in the same department. Further directions are offered by studying how materials from this corpus can be used to let students themselves explore the language used in a specific sub-corpus (in their specific rhetorical and argumentative structures), before they actually learn how to use it in guided and more independent activities. While hoping to be able to develop many of these ideas in future research, we hope to have offered here an adequate account of how the corpus was created, together with a few possible reasons for the potential interest we see in the dataset.

Acknowledgements: The authors wish to thank the members of CAP (*Comunicazione Accademica e Professionale*; Academic and Professional Communication).

Marina Bondi *is Full Professor of English Linguistics at the University of Modena and Reggio Emilia (Italy) and Founding Director of the CLAVIER centre (Corpus and LAnguage Variation In English Research). Her research centres on textual, pragmatic and phraseological aspects of academic and professional discourse across genres, discourse identities and media.*

Matteo Di Cristofaro *is Lecturer in Digital Humanities and Corpus Linguistics at the University of Modena and Reggio Emilia, and Researcher Fellow in Corpus Linguistics at the University of Pisa. His research focuses on interdisciplinary applied corpus linguistics, and on the implications that digital technologies have on the use and analysis of natural language.*

Works cited

Aguilar-Pérez, Marta and Sarah Khan. “Metadiscourse Use When Shifting from L1 to EMI Lecturing: Implications for Teacher Training.” *Innovation in Language Learning and Teaching* 16.4-5 (2022): 297-311.

Becker, Maria, Michael Bender and Marcus Müller. “Classifying Heuristic Textual Practices in Academic Discourse: A Deep Learning Approach to Pragmatics.” *International Journal of Corpus Linguistics* 25.4 (2020): 426-460.

Bier, Ada. “From Effective Lecturing Behaviour to Hidden Cognitions: A Preliminary Model Explaining the Language-Teaching Methodology Interface.” *Innovation in Language Learning and Teaching* 16.4-5 (2022): 351-365.

Bondi, Marina. “Changing Voices: Authorial Voice in Abstracts.” *Abstracts in Academic Discourse.* Edited by Marina Bondi and Rosa Lorés Sanz. New York: Peter Lang, 2014. 243-269.

---. “The Discourse Function of Contrastive Connectors in Academic Abstracts.” *Discourse Patterns in Spoken and Written Corpora.* Edited by Karin Aijmer and Anna-Brita Stenström. Amsterdam: John Benjamins, 2004. 139-156.

Boullosa, Beto, et al. “Integrating Knowledge-Supported Search into the INCEpTION Annotation Platform.” *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing (2018): 127-132.

Campagna, Sandra and Virginia Pulcini. "English as a Medium of Instruction in Italian Universities: Linguistic Policies, Pedagogical Implications." *Textus, English Studies in Italy* 1 (2014): 173-190.

Chambers, Angela. "What is Data-Driven Learning?" *The Routledge Handbook of Corpus Linguistics*. Edited by Anne O'Keeffe and Michael McCarthy. Milton Park: Routledge, 2012. 345-358.

Charles, Maggie, Diane Pecorari and Susan Hunston, edited by. *Academic Writing: At the Interface of Corpus and Discourse*. New York: Continuum, 2009.

Christ, Oliver. "A Modular and Flexible Architecture for an Integrated Corpus Query System." *Proceedings of COMPLEX '94* (1994): 1-10.

Corcoran, James N., Karen Englander and Laura-Mihaela Muresan, edited by. *Pedagogies and Policies for Publishing Research in English: Local Initiatives Supporting International Scholars*. New York: Routledge, 2019.

Costa, Francesca and James A. Coleman. "A Survey of English-Medium Instruction in Italian Higher Education." *International Journal of Bilingual Education and Bilingualism* 16.1 (2013): 3-19.

Costa, Francesca and Olivia Mair. "Multimodality and Pronunciation in ICLHE (Integration of Content and Language in Higher Education) Training." *Innovation in Language Learning and Teaching* 16.4-5 (2022): 281-296.

de Castilho, Richard Eckart, et al. "INCEpTION - Corpus-Based Data Science from Scratch." *Digital Infrastructures for Research (DI4R) 2018* (2018).

Dearden, Julie. *English as a Medium of Instruction – A Growing Global Phenomenon*. British Council, 2014.

Doiz, Aintzane, et al. *English-Medium Instruction at Universities: Global Challenges*. Bristol: Multilingual Matters, 2012.

Doiz, Aintzane and David Lasagabaster. "Looking into English-Medium Instruction Teachers' Metadiscourse: An ELF Perspective." *System* 105 (2022): 1-12.

Evert, Stefan and Andrew Hardie. "Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium." *International Journal of Corpus Linguistics* 17.3 (2011): 380-409.

Ferguson, Gibson, Carmen Pérez-Llantada and Ramòn Plo. "English as an International Language of Scientific Publication: A Study of Attitudes." *World Englishes* 30.1 (2011): 41-59.

Flowerdew, John. “Attitudes of Journal Editors to Nonnative Speaker Contributions.” *TESOL Quarterly* 35.1 (2001): 121-150.

---. “Writing for Scholarly Publication in English: The Case of Hong Kong.” *Journal of Second Language Writing* 8.2 (1999): 123-145.

Flowerdew, John and Pejman Habibie. *Introducing English for Research Publication Purposes*. London: Routledge, 2022.

Flowerdew, Lynne. “Applying Corpus Linguistics to Pedagogy: A Critical Evaluation.” *International Journal of Corpus Linguistics* 14.3 (2009): 393-417.

---. “Corpora for Eap Writing.” *The Routledge Handbook of Corpora and English Language Teaching and Learning.* Edited by Reka R. Jablonkai and Eniko Csomay. London: Routledge, 2022. 234-247.

---. “Corpus-Based Research and Pedagogy in EAP: From Lexis to Genre.” *Language Teaching* 48.1 (2015): 99-116.

Fortanet-Gómez, Inmaculada. *Towards a Multilingual Language Policy*. Bristol: Multilingual Matters, 2013.

Greenbaum, Sidney. *The Oxford English Grammar*. Oxford: Oxford University Press, 1996.

Hardie, Andrew. “CQPweb — Combining Power, Flexibility and Usability in a Corpus Analysis Tool.” *International Journal of Corpus Linguistics* 17.3 (2012): 380-409.

---. “Modest XML for Corpora: Not a Standard, but a Suggestion.” *ICAME Journal* 38.1 (2014): 73-103.

Hyland, Ken and Feng (Kevin) Jiang. *Academic Discourse and Global Publishing: Disciplinary Persuasion in Changing Times*. London: Routledge, 2019.

Hyland, Ken and Marina Bondi, edited by. *Academic Discourse across Disciplines*. New York: Peter Lang, 2006.

Hynninen, Niina. *Language Regulation in English as a Lingua Franca: Focus on Academic Spoken Discourse*. Berlin: De Gruyter Mouton, 2016.

Hynninen, Niina and Maria Kuteeva. “‘Good’ and ‘Acceptable’ English in L2 Research Writing: Ideals and Realities in History and Computer Science.” *Journal of English for Academic Purposes* 30 (2017): 53-65.

Jensen, Christian and Jacob Thøgersen. “Comprehensibility, Lecture Recall and Attitudes in EMI.” *Journal of English for Academic Purposes* 48 (2020): 1-12.

Kachru, Braj B., Yamuna Kachru and Cecil L. Nelson. *The Handbook of World Englishes*. Oxford: Blackwell Publishing, 2006.

Kielipankki. “The Korp Corpus Input Format.” *Kielipankki The Language Bank of Finland*

(2021). https://www.kielipankki.fi/development/korp/corpus-input-format/.

Kilgarriff, Adam, et al. "The Sketch Engine." *Proceedings of the 11th EURALEX International Congress* (2004). 105-116.

Kilgarriff, Adam, et al. "The Sketch Engine: Ten Years On." *Lexicography* 1.1 (2014): 7-36.

Klie, Jan-Christoph. "INCEpTION: Interactive Machine-Assisted Annotation." *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, 2018. 105.

Klie, Jan-Christoph, et al. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, 2018. 5-9.

Lillis, Theresa and Mary Jane Curry. *Academic Writing in a Global Context: The Politics and Practices of Publishing in English*. London: Routledge, 2010.

Mauranen, Anna. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press, 2012.

Mauranen, Anna, Carmen Pérez-Llantada and John M. Swales. "Academic Englishes: A Standardised Knowledge?" *The Routledge Handbook of World Englishes*, 2nd ed. London: Routledge, 2020. 659-676.

Mauranen, Anna, Niina Hynninen and Elina Ranta. "English as an Academic Lingua Franca: The ELFA Project." *English for Specific Purposes* 29.3 (2010): 183-190.

---. "Second Language Acquisition, World Englishes, and English as a Lingua Franca (ELF)." *World Englishes* 37.1 (2018): 106-119.

Mur-Dueñas, Pilar and Jolanta Šinkūnienė, edited by. *Intercultural Perspectives on Research Writing*. Amsterdam: John Benjamins, 2018.

---. "Self-Reference in Research Articles across Europe and Asia: A Review of Studies." *Brno Studies in English* 1 (2016): 71-92.

Paré, Anthony, Doreen Starke-Meyerring and Lynn McAlpine. "Knowledge and Identity Work in the Supervision of Doctoral Student Writing: Shaping Rhetorical Subjects." *Writing in Knowledge Societies.* Edited by Doreen Starke-Meyerring, et al. Anderson: Parlor Press, 2011. 215-236.

Pérez-Llantada, Carmen. "Formulaic Language in L1 and L2 Expert Academic Writing: Convergent and Divergent Usage." *Journal of English for Academic Purposes* 14 (2014): 84-94.

Pérez-Llantada, Carmen, Ramón Plo and Gibson R. Ferguson. ""You Don't Say What You Know, Only What You Can": The Perceptions and Practices of Senior Spanish Academics Regarding Research Dissemination in English." *English for Specific Purposes* 30.1 (2011): 18-30.

Qi, Peng, et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020. 101-108.

Römer, Ute and Ann Arbor. "English in Academia: Does Nativeness Matter?" *Anglistik: International Journal of English Studies* 20.2 (2009): 89-100.

Rozycki, William and Neil H. Johnson. "Non-Canonical Grammar in Best Paper Award Winners in Engineering." *English for Specific Purposes* 32.3 (2013): 157-169.

Sano, Hikomaro. "The World's Lingua Franca of Science." *English Today* 18.4 (2002): 45-49.

Suresh, Canagarajah A. "'Nondiscursive' Requirements in Academic Publishing, Material Resources of Periphery Scholars, and the Politics of Knowledge Production." *Written Communication* 13.4 (1996): 435-472.

Tribble, Christopher. "Corpora and Corpus Analysis: New Windows on Academic Writing." *Academic Discourse.* Edited by John Flowerdew. London: Routledge, 2002.

---. "ELFA vs. Genre: A New Paradigm War in EAP Writing Instruction?" *Journal of English for Academic Purposes* 25 (2017): 30-44.

Wu, Xue, Anna Mauranen and Lei Lei. "Syntactic Complexity in English as a Lingua Franca Academic Writing." *Journal of English for Academic Purposes* 43 (2020): 1-13.